(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 242 589 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
13.11.2013 Bulletin 2013/46

(21) Application number: **00989578.0**

(22) Date of filing: **29.12.2000**

(51) Int Cl.:
C12N 9/78 *(2006.01)* ***C12P 13/04*** *(2006.01)*
C12P 7/42 *(2006.01)* ***C12P 41/00*** *(2006.01)*

(86) International application number:
PCT/US2000/035555

(87) International publication number:
WO 2001/048175 (05.07.2001 Gazette 2001/27)

(54) **METHODS FOR PRODUCING ENANTIOMERICALLY PURE ALPHA-SUBSTITUTED CARBOXYLIC ACIDS**

METHODEN ZUR HERSTELLUNG ENANTIOMERENREINER ALPHA-SUBSTITUIERTER CARBONSÄUREN

METHODES DESTINEES A LA PRODUCTION D'ACIDES CARBOXYLIQUES ALPHA-SUBSTITUES ENANTIOMERIQUEMENT PURS

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR

(30) Priority: **29.12.1999 US 173609 P**
07.12.2000 US 254414 P

(43) Date of publication of application:
25.09.2002 Bulletin 2002/39

(60) Divisional application:
10177452.9 / 2 258 840

(73) Proprietor: **Verenium Corporation**
San Diego, CA 92121 (US)

(72) Inventors:
- **MADDEN, Mark**
  deceased (US)
- **WEINER, David, Paul**
  Del Mar, CA 92014 (US)
- **CHAPLIN, Jennifer, Ann**
  San Diego, CA 92130 (US)
- **Dr. Dan Robertson**
  California 92075 (CA)

(74) Representative: **Wakerley, Helen Rachael**
Reddie & Grose LLP
16 Theobalds Road
London WC1X 8PL (GB)

(56) References cited:
EP-A- 0 332 379 **EP-A- 0 449 648**
EP-A- 0 502 476 **EP-A- 0 610 048**
EP-A- 0 610 049 **EP-A- 0 711 836**
EP-A- 0 780 471 **WO-A-86/07386**

- **DATABASE WPI Section Ch, Week 199217 Derwent Publications Ltd., London, GB; Class B05, AN 1992-137926 XP002165677 & JP 04 079894 A (NIPPON MINING CO), 13 March 1992 (1992-03-13)**
- **DATABASE BIOSIS [Online] BIOSCIENCES INFORMATION SERVICE, PHILADELPHIA, PA, US; 1991 TAILLADES J ET AL: "ENZYMATIC HYDROLYSIS OF RACEMIC PHENYLALANINAMIDE WITH PRONASE IMMOBILIZED ON KETONIC POLYMER" Database accession no. PREV199192083701 XP002165676 & BULLETIN DE LA SOCIETE CHIMIQUE DE FRANCE, vol. 128, no. 3, 1991, pages 423-430, ISSN: 0037-8968**
- **WIYAKRUTTA SUTHEP ET AL: "A stereo-inverting D-phenylglycine aminotransferase from Pseudomonas stutzeri ST-201: Purification, characterization and application for D-phenylglycine synthesis." JOURNAL OF BIOTECHNOLOGY, vol. 55, no. 3, 1997, pages 193-203, XP002165675 ISSN: 0168-1656**
- **BORK P ET AL: "NEW FAMILY OF CARBON-NITROGEN HYDROLASES" PROTEIN SCIENCE, CAMBRIDGE UNIVERSITY PRESS, CAMBRIDGE, GB, vol. 3, no. 8, 1994, pages 1344-1346, XP000924973 ISSN: 0961-8368**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Printed by Jouve, 75001 PARIS (FR)

(Cont. next page)

- **LEVY-SCHIL S ET AL: "Aliphatic nitrilase from a soil-isolated Comamonas testosteroni sp.: gene cloning and overexpression, purification and primary structure" GENE, ELSEVIER BIOMEDICAL PRESS. AMSTERDAM, NL, vol. 161, no. 1, 8 August 1995 (1995-08-08), pages 15-20, XP004042111 ISSN: 0378-1119**
- **NOVO C ET AL: "PSEUDOMANAS AERUGINOSA ALIPHATIC AMIDASE IS RELATED TO THE NITRILASE/CYANIDE HYDRATASE ENZYME FAMILY AND CYS 166 IS PREDICTED TO BE THE ACTIVE SITE NECLEOPHILE OF THE CATALYTIC MECHANISM" FEBS LETTERS, ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM, NL, vol. 367, no. 3, 1995, pages 275-279, XP000915286 ISSN: 0014-5793**
- **WATANABE A ET AL: "CLONING AND EXPRESSION OF A GENE ENCODING CYANIDASE FROM PSEUDOMONAS STUTZERI AK61" APPLIED MICROBIOLOGY AND BIOTECHNOLOGY, SPRINGER VERLAG, BERLIN, DE, vol. 50, no. 1, July 1998 (1998-07), pages 93-97, XP000924971 ISSN: 0175-7598**
- **DURAN R ET AL: "CHARACTERIZATION OF NITRILE HYDRATASE GENES CLONED BY DNA SCREEING FROM RHODOCOCCUS ERYTHROPOLIS" BIOSCIENCE BIOTECHNOLOGY BIOCHEMISTRY, JAPAN SOC. FOR BIOSCIENCE, BIOTECHNOLOGY AND AGROCHEM. TOKYO, JP, vol. 57, no. 8, 1993, pages 1323-1328, XP002050202 ISSN: 0916-8451**
- **KOMEDA H ET AL: "TRANSCRIPTIONAL REGULATION OF THE RHODOCOCCUS RHODOCHROUS J1 NITA GENE ENCODING A NITRILASE" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, NATIONAL ACADEMY OF SCIENCE. WASHINGTON, US, vol. 93, no. 20, 1 October 1996 (1996-10-01), pages 10572-10577, XP002045089 ISSN: 0027-8424**
- **DESANTIS GRACE ET AL: "An enzyme library approach to biocatalysis: Development of nitrilases for enantioselective production of carboxylic acid derivatives", JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, vol. 124, no. 31, 7 August 2002 (2002-08-07), pages 9024-9025, ISSN: 0002-7863**

Remarks:

The file contains technical information submitted after the application was filed and not included in this specification

Description

FIELD OF THE INVENTION

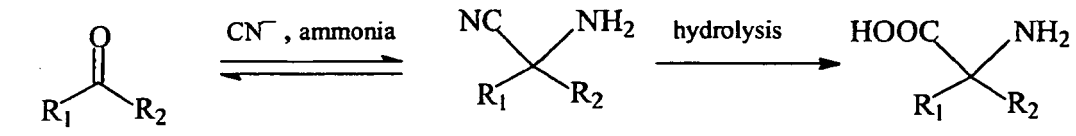
[0001] The present invention relates generally to methods for producing enantiomerically pure α-substituted carboxylic acids. The invention relates more particularly to methods for producing enantiomerically pure α-amino acids in a single reaction vessel comprising the use of a stereoselective nitrilase in the presence of a cyanide compound, an ammonia or amine compound, and an aldehyde or a ketone.

BACKGROUND OF THE INVENTION

[0002] The continuing importance of asymmetric organic synthesis in drug design and discovery has fueled the search for new synthetic methods and chiral precursors. This research effort has resulted in the identification of many new synthetic methods and chiral precursors which have been utilized in developing complex molecules of biological interest.

[0003] One important class of chiral molecules are the α-substituted carboxylic acids. These molecules have long been recognized as important chiral precursors to a wide variety of complex biologically active molecules. In particular, a great deal of research effort has been dedicated to the development of methods for the synthesis of enantiomerically pure α-amino acids. Recently, there has been an increasing demand for enantiomerically pure α-amino acids for a variety of uses, including, for example, chiral medicines.

[0004] A common synthetic route to α-amino acids is the Strecker synthesis, shown below:



aldehyde or ketone

Reversible addition of cyanide and ammonia to the aldehyde or ketone produces an amino nitrile intermediate, which, upon hydrolysis, yields the desired α-amino acid. Although this synthesis has been used to produce racemic amino acids on an industrial scale, there has been only moderate success in developing chiral versions of the Strecker synthesis.

Acknowledgment of Cited Art

[0005] WO 86/07386 discloses a method of preparing optically active amino acids or amino acid amides by converting an amino nitrile using an enantioselective nitrilase.

[0006] In EP 0610048 microbial cells harvested from Gordona terrae are used to produce optically active α-hydrocarboxylic acid having a phenyl group.

[0007] Aureobacterium testaceum provides a source of cells used in the production of R(-)-mandelic acid derivatives in EP 0449648.

[0008] EP 0610049 discloses the use of microbial cells from Rhodococcus sp or Brevibacterium acetylicum in the synthesis of L-3-phenyllactic acid.

[0009] EP 0780471 discloses a cloned nitrilase gene from Gordona terrae MA-1.

[0010] The study of nitrilases, looking for sequence similarities is disclosed, for example, in Protein Science (1994), 3: 1334-1348 "New family of carbon-nitrogen hydrolases".

[0011] Gene 161 (1995) 15 20 discloses the identification and purification of a nitrilase from Comamonas testosteroni sp and its alignment with known nitrilases.

[0012] Accordingly, there is a need in the art for efficient, inexpensive, high-yield synthetic methods for producing enantiomerically pure α-substituted carboxylic acids, such as, for example, α-amino acids and α-hydroxy acids.

SUMMARY OF THE INVENTION

[0013] The invention provides nitrilase polypeptides and nucleic acid sequences encoding such nitrilase polypeptides. In one aspect, the invention provides a substantially purified isolated or recombinant polypeptide having an amino acid sequence as set forth in SEQ ID NO:4 and sequences having at least 90% identity thereto and having nitrilase activity.

[0014] In another aspect, the invention provides an isolated or recombinant nucleic acid sequence encoding an amino acid sequence as set forth in SEQ ID NO:4 and sequences having at least 90% identity thereto and having nitrilase activity. As set out in claim 2, the sequence may have at least 90% identity with SEQ ID NO:3.

[0015] In accordance with the present invention there are further provided methods for producing enantiomerically pure α-substituted carboxylic acids, such as, for example, α-amino acids and α-hydroxy acids. The methods include combining an aldehyde or ketone with a cyanide and an ammonia-containing compound or an ammonium salt, in the presence of a nitrilase which stereoselectively hydrolyzes the amino nitrile or cyanohydrin intermediate, under conditions sufficient to produce the carboxylic acid. The nitrilase comprises a polypeptide of the invention or is encoded by a nucleic acid sequence of the invention.

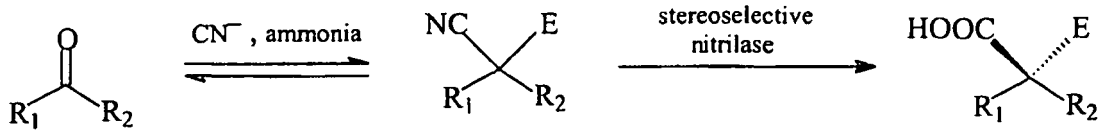
[0016] The invention provides a method for producing an enantiomerically pure α-substituted carboxylic acid. The method includes contacting an aldehyde or ketone with a cyanide containing compound and an ammonia-containing compound or an ammonium salt or an amine, and stereoselectively hydrolyzing the resulting amino nitrile or cyanohydrin intermediate with the said nitrilase or a polypeptide having nitrilase activity, wherein the nitrilase is sufficiently active to perform the hydrolysis in the presence of the reaction components, under conditions and for a time sufficient to produce the carboxylic acid.

[0017] Aspects of the invention are set out in the accompanying claims.

DETAILED DESCRIPTION OF THE INVENTION

[0018] In accordance with the present invention, there are provided methods for producing enantiomerically pure α-substituted carboxylic acids. The methods of the invention include contacting an aldehyde or ketone with a cyanide-containing compound, preferably a metal or gaseous cyanide compound, and an ammonia-containing compound or an ammonium salt or an amine, and stereoselectively hydrolyzing the resulting amino nitrile or cyanohydrin intermediate with a nitrilase of the invention, wherein the nitrilase is sufficiently active to perform the hydrolysis in the presence of cyanide and ammonia. This stereoselective synthesis is outlined in Scheme 1.

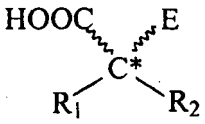
Scheme 1



aldehyde or ketone

[0019] Since the formation of the amino nitrile or cyanohydrin is a reversible process, the use of a stereoselective nitrilase provides chiral induction, thereby producing the desired amino acid or hydroxy acid enantiomer in 100% theoretical yield. Indeed, even aryl aldehydes and ketones (wherein the equilibrium disfavors amino nitrile or cyanohydrin formation) are effectively converted to chiral α-carboxylic acids. Moreover, since the stereoselective nitrilases contemplated for use in the practice of the present invention are able to perform the hydrolysis in the presence of cyanide and ammonia, invention methods provide the additional advantage of producing the desired enantiomerically pure α-substituted carboxylic acids in a single reaction vessel. The nitrilase is SEQ ID NO:4, or a polypeptide of the first aspect of the invention.

[0020] The enantiomerically pure α-substituted carboxylic acids produced by the methods of the present invention have the following structure:



wherein:

R_1 and R_2 are each independently -H, substituted or unsubstituted alkyl, alkenyl, alkynyl, aryl, heteroaryl, cycloalkyl, heterocyclic, wherein said substituents are lower alkyl, hydroxy, alkoxy, mercapto, cycloalkyl, heterocyclic, aryl, heteroaryl, aryloxy, or halogen or optionally R_1 and R_2 are linked to cooperate to form a functional cyclic moiety, and E is $-N(R_x)_2$ or -OH, wherein each R_x is -H or lower alkyl.

[0021] As used herein, the term "alkyl" refers to a monovalent straight or branched chain or cyclic radical of from one to twenty-four carbon atoms, including methyl, ethyl, n-propyl, isopropyl, n-butyl, isobutyl, tert-butyl, n-hexyl, and the like. The term "lower alkyl" refers to a monovalent straight or branched chain or cyclic radical of from one to about six

carbon atoms.

[0022] As used herein, "substituted alkyl" comprises alkyl groups further bearing one or more substituents selected from hydroxy, alkoxy (of a lower alkyl group), mercapto (of a lower alkyl group), cycloalkyl, substituted cycloalkyl, heterocyclic, substituted heterocyclic, aryl, substituted aryl, heteroaryl, substituted heteroaryl, aryloxy, substituted aryloxy, halogen, trifluoromethyl, cyano, nitro, nitrone, amino, amido, -C(O)H, acyl, oxyacyl, carboxyl, carbamate, sulfonyl, sulfonamide, sulfuryl, and the like.

[0023] As used herein, "alkenyl" refers to straight or branched chain hydrocarbyl groups having one or more carbon-carbon double bonds, and having in the range of about 2 up to 24 carbon atoms, and "substituted alkenyl" refers to alkenyl groups further bearing one or more substituents as set forth above.

[0024] As used herein, "cycloalkyl" refers to cyclic ring-containing groups containing in the range of about 3 up to 8 carbon atoms, and "substituted cycloalkyl" refers to cycloalkyl groups further bearing one or more substituents as set forth above.

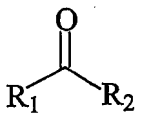
[0025] As used herein, "heterocyclic" refers to cyclic (i.e., ring-containing) groups containing one or more heteroatoms (e.g., N, O, S, or the like) as part of the ring structure, and having in the range of 3 up to 14 carbon atoms and "substituted heterocyclic" refers to heterocyclic groups further bearing one or more substituents as set forth above.

[0026] As used herein, "alkynyl" refers to straight or branched chain hydrocarbyl groups having at least one carbon-carbon triple bond, and having in the range of about 2 up to 12 carbon atoms, and "substituted alkynyl" refers to alkynylene groups further bearing one or more substituents as set forth above.

[0027] As used herein, "aryl" refers to aromatic groups having in the range of 6 up to 14 carbon atoms and "substituted aryl" refers to aryl groups further bearing one or more substituents as set forth above.

[0028] In preferred embodiments, the enantiomerically pure α-substituted carboxylic acid produced by the methods of the present invention is an α-amino acid or α-hydroxy acid. In particularly preferred embodiments the enantiomerically pure α-amino acid is D-phenylalanine, D-phenylglycine, L-methylphenylglycine, L-tert-leucine, D-alanine, or D-hydroxynorleucine, R-pantolactone, 2-chloromandelic acid, (S)-and (R)-mandelic acid and the enantiomerically pure α-hydroxy acid is (S)-cyclohexylmandelic acid.

[0029] Aldehydes and ketones contemplated for use in the practice of the present invention have the following structure:



wherein R_1 and R_2 are as defined above. In preferred embodiments, at least one of R_1 and R_2 is an aryl group.

[0030] Metal cyanides contemplated for use in the practice of the present invention include the alkali metal cyanides LiCN, NaCN, KCN, RbCN, and CsCn. Preferred alkali metal cyanides include LiCN, NaCN, and KCN. A most preferred alkali metal cyanide is KCN. In addition, gaseous cyanides are useful in the method of the invention.

[0031] Ammonia or an ammonium salt or an amine may be used in accordance with the present invention. Ammonium salts contemplated for use in the practice of the present invention have the formula $NH_2(R)_2{}^+X^-$, wherein each R is independently -H or lower alkyl, and X is fluoride, chloride, bromide, or iodide or any counter ion. Thus, when R is lower alkyl, the methods of the present invention also provide *N*-substituted or *N*,*N*-disubstituted enantiomerically pure α-amino acids. In a preferred embodiment, the ammonium salt is $NH_4{}^+Cl^-$.

[0032] Nitrilases contemplated for use in the practice of the present invention include those which are sufficiently robust to stereoselectively hydrolyze the transient amino nitrile or cyanohydrin under Strecker conditions, i.e., in the presence of cyanide and ammonia. Such nitrilases include, for example, those set forth in SEQ ID No: 4.

[0033] The phrases "nucleic acid" or "nucleic acid sequence" as used herein refer to an oligonucleotide, nucleotide, polynucleotide, or to a fragment of any of these, to DNA or RNA of genomic or synthetic origin which may be single-stranded or double-stranded and may represent a sense or antisense strand, peptide nucleic acid (PNA), or to any DNA-like or RNA-like material, natural or synthetic in origin. In one embodiment, a "nucleic acid sequence" of the invention includes, for example, a sequence encoding a polypeptide as set forth in SEQ ID No: 4, and variants thereof. In another embodiment, a "nucleic acid sequence" of the invention includes, for example, a sequence as set forth in SEQ ID No: 3, sequences complemetary thereto, and variants thereof.

[0034] A "coding sequence" or a "nucleotide sequence encoding" a particular polypeptide or protein, is a nucleic acid sequence which is transcribed and translated into a polypeptide or protein when placed under the control of appropriate regulatory sequences.

[0035] The term "gene" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as, where applicable, intervening sequences (introns) between individual coding segments (exons).

[0036] "Amino acid" or "amino acid sequence" as used herein refer to an oligopeptide, peptide, polypeptide, or protein

sequence, or to a fragment, portion, or subunit of any of these, and to naturally occurring or synthetic molecules. In one embodiment, an "amino acid sequence" or "polypeptide sequence" of the invention includes, for example, a sequence as set forth in SEQ ID No: 4, and variants thereof. In another embodiment, an "amino acid sequence" of the invention includes, for example, a sequence encoded by a polynucleotide having a sequence as set forth in SEQ ID No: 3, sequences complemetary thereto and variants thereof.

[0037] The term "polypeptide" as used herein, refers to amino acids joined to each other by peptide bonds or modified peptide bonds, i.e., peptide isosteres, and may contain modified amino acids other than the 20 gene-encoded amino acids. The polypeptides may be modified by either natural processes, such as post-translational processing, or by chemical modification techniques which are well known in the art. Modifications can occur anywhere in the polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given polypeptide. Also a given polypeptide may have many types of modifications. Modifications include acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of a phosphytidylinositol, cross-linking cyclization, disulfide bond formation, demethylation, formation of covalent cross-links, formation of cysteine, formation of pyroglutamate, formylation, gamma-carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristolyation, oxidation, pergylation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, and transfer-RNA mediated addition of amino acids to protein such as arginylation. (See Proteins - Structure and Molecular Properties 2nd Ed., T.E. Creighton, W.H. Freeman and Company, New York (1993); Posttranslational Covalent Modification of Proteins, B.C. Johnson, Ed., Academic Press, New York, pp. 1-12 (1983)).

[0038] As used herein, the term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

[0039] As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual nucleic acids obtained from a library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The purified nucleic acids of the invention have been purified from the remainder of the genomic DNA in the organism by at least 10^4-10^6 fold. However, the term "purified" also includes nucleic acids which have been purified from the remainder of the genomic DNA or from other sequences in a library or other environment by at least one order of magnitude, typically two or three orders, and more typically four or five orders of magnitude.

[0040] As used herein, the term "recombinant" means that the nucleic acid is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the nucleic acids will represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. Typically, the enriched nucleic acids represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More typically, the enriched nucleic acids represent 50% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. In a one embodiment, the enriched nucleic acids represent 90% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. "Recombinant" polypeptides or proteins refer to polypeptides or proteins produced by recombinant DNA techniques; i.e., produced from cells transformed by an exogenous DNA construct encoding the desired polypeptide or protein. "Synthetic" polypeptides or protein are those prepared by chemical synthesis. Solid-phase chemical peptide synthesis methods can also be used to synthesize the polypeptide or fragments of the invention. Such method have been known in the art since the early 1960's (Merrifield, R. B., J. Am. Chem. Soc., 85:2149-2154,1963) (See also Stewart, J. M. and Young, J. D., Solid Phase Peptide Synthesis, 2 ed., Pierce Chemical Co., Rockford, Ill., pp. 11-12)) and have recently been employed in commercially available laboratory peptide design and synthesis kits (Cambridge Research Biochemicals). Such commercially available laboratory kits have generally utilized the teachings of H. M. Geysen et al, Proc. Natl. Acad. Sci., USA, 81:3998 (1984) and provide for synthesizing peptides upon the tips of a multitude of "rods" or "pins" all of which are connected to a single plate. When such a system is utilized, a plate of rods or pins is inverted and inserted into a second plate of corresponding wells or reservoirs, which contain solutions for attaching or anchoring an appropriate amino acid to the pin's or rod's tips. By repeating such a process step, i.e., inverting and inserting the rod's and pin's tips into appropriate solutions, amino acids are built into desired peptides. In addition, a number of available FMOC peptide synthesis systems are available. For example, assembly of a polypeptide or fragment can be carried out on a solid support using an Applied Biosystems, Inc. Model 431A automated peptide synthesizer. Such equipment

provides ready access to the peptides of the invention, either by direct synthesis or by synthesis of a series of fragments that can be coupled using other known techniques.

[0041] A promoter sequence is "operably linked to" a coding sequence when RNA polymerase which initiates transcription at the promoter will transcribe the coding sequence into mRNA.

[0042] "Plasmids" are designated by a lower case p preceded and/or followed by capital letters and/or numbers. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accord with published procedures. In addition, equivalent plasmids to those described herein are known in the art and will be apparent to the ordinarily skilled artisan.

[0043] "Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 g of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 1 of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 g of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37□C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the gel electrophoresis may be performed to isolate the desired fragment.

[0044] "Oligonucleotide" refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides have no 5' phosphate and thus will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated.

[0045] The phrase "substantially identical" in the context of two nucleic acid sequences or polypeptides, refers to two or more sequences that have at least 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the known sequence comparison algorithms or by visual inspection. Typically, the substantial identity exists over a region of at least about 100 residues, and most commonly the sequences are substantially identical over at least about 150-200 residues. In some embodiments, the sequences are substantially identical over the entire length of the coding regions.

[0046] Additionally a "substantially identical" amino acid sequence is a sequence that differs from a reference sequence by one or more conservative or non-conservative amino acid substitutions, deletions, or insertions, particularly when such a substitution occurs at a site that is not the active site of the molecule, and provided that the polypeptide essentially retains its functional properties. A conservative amino acid substitution, for example, substitutes one amino acid for another of the same class (e.g., substitution of one hydrophobic amino acid, such as isoleucin, valine, leucine, or methionine, for another, or substitution of one polar amino acid for another, such as substitution of arginine for lysine, glutamic acid for aspartic acid or glutamine for asparagine). One or more amino acids can be deleted, for example, from a haloalkane dehalogenase polypeptide, resulting in modification of the structure of the polypeptide, without significantly altering its biological activity. For example, amino- or carboxyl-terminal amino acids that are not required for haloalkane dehalogenase biological activity can be removed. Modified polypeptide sequences of the invention can be assayed for haloalkane dehalogenase biological activity by any number of methods, including contacting the modified polypeptide sequence with an haloalkane dehalogenase substrate and determining whether the modified polypeptide decreases the amount of specific substrate in the assay or increases the bioproducts of the enzymatic reaction of a functional haloalkane dehalogenase polypeptide with the substrate.

[0047] "Fragments" as used herein are a portion of a naturally occurring or recombinant protein which can exist in at least two different conformations. Fragments can have the same or substantially the same amino acid sequence as the naturally occurring protein. "Substantially the same" means that an amino acid sequence is largely, but not entirely, the same, but retains at least one functional activity of the sequence to which it is related. In general two amino acid sequences are "substantially the same" or "substantially homologous" if they are at least about 70, but more typically about 85% or more identical. Fragments which have different three dimensional structures as the naturally occurring protein are also included. An example of this, is a "pro-form" molecule, such as a low activity proprotein that can be modified by cleavage to produce a mature enzyme with significantly higher activity.

[0048] "Hybridization" refers to the process by which a nucleic acid strand joins with a complementary strand through base pairing. Hybridization reactions can be sensitive and selective so that a particular sequence of interest can be identified even in samples in which it is present at low concentrations. Suitably stringent conditions can be defined by, for example, the concentrations of salt or formamide in the prehybridization and hybridization solutions, or by the hybridization temperature, and are well known in the art. In particular, stringency can be increased by reducing the concentration of salt, increasing the concentration of formamide, or raising the hybridization temperature.

[0049] For example, hybridization under high stringency conditions could occur in about 50% formamide at about 37°C to 42°C. Hybridization could occur under reduced stringency conditions in about 35% to 25% formamide at about 30°C to 35°C. In particular, hybridization could occur under high stringency conditions at 42°C in 50% formamide, 5X SSPE,

0.3% SDS, and 200 n/ml sheared and denatured salmon sperm DNA. Hybridization could occur under reduced stringency conditions as described above, but in 35% formamide at a reduced temperature of 35°C. The temperature range corresponding to a particular level of stringency can be further narrowed by calculating the purine to pyrimidine ratio of the nucleic acid of interest and adjusting the temperature accordingly. Variations on the above ranges and conditions are well known in the art.

[0050] The term "variant" refers to polynucleotides or polypeptides of the invention modified at one or more base pairs, codons, introns, exons, or amino acid residues (respectively) yet still retain the biological activity of the nitrilase of the invention. Variants can be produced by any number of means including methods such as, for example, error-prone PCR, shuffling, oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, in vivo mutagenesis, cassette mutagenesis, recursive ensemble mutagenesis, exponential ensemble mutagenesis, site-specific mutagenesis, ligation reassembly, GSSM and any combination thereof.

[0051] Polynucleotides as hereinabove described can be introduced into a suitable host cell. A suitable host cell is any cell which is capable of promoting recombination and/or reductive reassortment. The selected polynucleotides are preferably already in a vector which includes appropriate control sequences. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or preferably, the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis et al., 1986).

[0052] As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Salmonella typhimurium; fungal cells, such as yeast; insect cells such as Drosophila S2 and Spodoptera Sf9; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; and plant cells. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

[0053] With particular references to various mammalian cell culture systems that can be employed to express recombinant protein, examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described in "SV40-transformed simian cells support the replication of early SV40 mutants" (Gluzman, 1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

[0054] Host cells containing the polynucleotides of interest can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan. The clones which are identified as having the specified enzyme activity may then be sequenced to identify the polynucleotide sequence encoding an enzyme having the enhanced activity.

[0055] As representative examples of expression vectors which may be used there may be mentioned viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral DNA (e.g., vaccinia, adenovirus, foul pox virus, pseudorabies and derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, aspergillus and yeast). Thus, for example, the DNA may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used so long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed with the present invention.

[0056] A preferred type of vector for use in the present invention contains an f-factor origin replication. The f-factor (or fertility factor) in E. coli is a plasmid which effects high frequency transfer of itself during conjugation and less frequent transfer of the bacterial chromosome itself. A particularly preferred embodiment is to use cloning vectors, referred to as "fosmids" or bacterial artificial chromosome (BAC) vectors. These are derived from E. coli f-factor which is able to stably integrate large segments of genomic DNA. When integrated with DNA from a mixed uncultured environmental sample, this makes it possible to achieve large genomic fragments in the form of a stable "environmental DNA library."

[0057] Another type of vector for use in the present invention is a cosmid vector. Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. Cloning into cosmid vectors is described in detail in "Molecular Cloning: A laboratory Manual" (Sambrook et al., 1989).

[0058] The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence (s) (promoter) to direct RNA synthesis. Particular named bacterial promoters include lad, lacZ, T3, T7, gpt, lambda PR, PL and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of

ordinary skill in the art. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or tetracycline or ampicillin resistance in E. coli.

[0059] One aspect of the invention is an isolated nucleic acid comprising SEQ ID No:3, sequences substantially identical thereto, sequences complementary thereto. The isolated, nucleic acids may comprise DNA, including cDNA, genomic DNA, and synthetic DNA. The DNA may be double-stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. Alternatively, the isolated nucleic acids may comprise RNA.

[0060] As discussed in more detail below, the isolated nucleic acid sequences of the invention may be used to prepare the polypeptide of SEQ ID No:4, and sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the polypeptides of SEQ ID No:4, and sequences substantially identical thereto.

[0061] Accordingly, another aspect of the invention is an isolated nucleic acid sequence which encodes one of the polypeptides of SEQ ID No:4, sequences substantially identical thereto. The coding sequences of these nucleic acids may be identical to the coding sequence of SEQ ID No:3, or a fragment thereof or may be different coding sequences which encode the polypeptide of SEQ ID No:4, sequences substantially identical thereto as a result of the redundancy or degeneracy of the genetic code. The genetic code is well known to those of skill in the art and can be obtained, for example, on page 214 of B. Lewin, Genes VI, Oxford University Press, 1997, the disclosure of which is incorporated herein by reference.

[0062] The isolated nucleic acid sequence which encodes the polypeptide of SEQ ID No:4, and sequences substantially identical thereto, may include, but is not limited to: only a coding sequence of SEQ ID No:3, and sequences substantially identical thereto, and additional coding sequences, such as leader sequences or proprotein sequences and non-coding sequences, such as introns or non-coding sequences 5' and/or 3' of the coding sequence. Thus, as used herein, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

[0063] Alternatively, the nucleic acid sequences of the invention may be mutagenized using conventional techniques, such as site directed mutagenesis, or other techniques familiar to those skilled in the art, to introduce silent changes into the polynucleotides of SEQ ID NO:3, and sequences substantially identical thereto. As used herein, "silent changes" include, for example, changes which do not alter the amino acid sequence encoded by the polynucleotide. Such changes may be desirable in order to increase the level of the polypeptide produced by host cells containing a vector encoding the polypeptide by introducing codons or codon pairs which occur frequently in the host organism.

[0064] The invention also relates to polynucleotides which have nucleotide changes which result in amino acid substitutions, additions, deletions, fusions and truncations in the polypeptide of the invention (SEQ ID No: 4). Such nucleotide changes may be introduced using techniques such as site directed mutagenesis, random chemical mutagenesis, exonuclease III deletion, and other recombinant DNA techniques. Alternatively, such nucleotide changes may be naturally occurring allelic variants which are isolated by identifying nucleic acid sequences which specifically hybridize to probes comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of the sequence of SEQ ID No: 3, and sequences substantially identical thereto (or the sequences complementary thereto) under conditions of high, moderate, or low stringency as provided herein.

[0065] The isolated nucleic acid of SEQ ID No: 3, sequences substantially identical thereto, complementary sequences, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of one of the foregoing sequences may also be used as probes to determine whether a biological sample, such as a soil sample, contains an organism having a nucleic acid sequence of the invention or an organism from which the nucleic acid was obtained. In such procedures, a biological sample potentially harboring the organism from which the nucleic acid was isolated is obtained and nucleic acids are obtained from the sample. The nucleic acids are contacted with the probe under conditions which permit the probe to specifically hybridize to any complementary sequences which are present therein.

[0066] Where necessary, conditions which permit the probe to specifically hybridize to complementary sequences may be determined by placing the probe in contact with complementary sequences from samples known to contain the complementary sequence as well as control sequences which do not contain the complementary sequence. Hybridization conditions, such as the salt concentration of the hybridization buffer, the formamide concentration of the hybridization buffer, or the hybridization temperature, may be varied to identify conditions which allow the probe to hybridize specifically to complementary nucleic acids.

[0067] If the sample contains the organism from which the nucleic acid was isolated, specific hybridization of the probe is then detected. Hybridization may be detected by labeling the probe with a detectable agent such as a radioactive isotope, a fluorescent dye or an enzyme capable of catalyzing the formation of a detectable product.

[0068] Many methods for using the labeled probes to detect the presence of complementary nucleic acids in a sample are familiar to those skilled in the art. These include Southern Blots, Northern Blots, colony hybridization procedures, and dot blots. Protocols for each of these procedures are provided in Ausubel et al. Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook et al., Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989.

[0069] Alternatively, more than one probe (at least one of which is capable of specifically hybridizing to any complementary sequences which are present in the nucleic acid sample), may be used in an amplification reaction to determine whether the sample contains an organism containing a nucleic acid sequence of the invention (e.g., an organism from which the nucleic acid was isolated). Typically, the probes comprise oligonucleotides. In one embodiment, the amplification reaction may comprise a PCR reaction. PCR protocols are described in Ausubel and Sambrook, supra. Alternatively, the amplification may comprise a ligase chain reaction, 3SR, or strand displacement reaction. (See Barany, F., "The Ligase Chain Reaction in a PCR World", PCR Methods and Applications 1:5-16, 1991; E. Fahy et al., "Self-sustained Sequence Replication (3SR): An Isothermal Transcription-based Amplification System Alternative to PCR", PCR Methods and Applications 1:25-33, 1991; and Walker G.T. et al., "Strand Displacement Amplification-an Isothermal in vitro DNA Amplification Technique", Nucleic Acid Research 20:1691-1696, 1992. In such procedures, the nucleic acids in the sample are contacted with the probes, the amplification reaction is performed, and any resulting amplification product is detected. The amplification product may be detected by performing gel electrophoresis on the reaction products and staining the gel with an intercalator such as ethidium bromide. Alternatively, one or more of the probes may be labeled with a radioactive isotope and the presence of a radioactive amplification product may be detected by autoradiography after gel electrophoresis.

[0070] Probes derived from sequences near the ends of a sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, may also be used in chromosome walking procedures to identify clones containing genomic sequences located adjacent to the nucleic acid sequences as set forth above. Such methods allow the isolation of genes which encode additional proteins from the host organism.

[0071] An isolated nucleic acid sequence as set forth in SEQ ID No: 3, sequences substantially identical thereto, sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300,400, or 500 consecutive bases of one of the foregoing sequences may be used as probes to identify and isolate related nucleic acids. In some embodiments, the related nucleic acids may be cDNAs or genomic DNAs from organisms other than the one from which the nucleic acid was isolated. For example, the other organisms may be related organisms. In such procedures, a nucleic acid sample is contacted with the probe under conditions which permit the probe to specifically hybridize to related sequences. Hybridization of the probe to nucleic acids from the related organism is then detected using any of the methods described above.

[0072] In nucleic acid hybridization reactions, the conditions used to achieve a particular level of stringency will vary, depending on the nature of the nucleic acids being hybridized. For example, the length, degree of complementarity, nucleotide sequence composition (e.g., GC v. AT content), and nucleic acid type (e.g., RNA v. DNA) of the hybridizing regions of the nucleic acids can be considered in selecting hybridization conditions. An additional consideration is whether one of the nucleic acids is immobilized, for example, on a filter.

[0073] Hybridization may be carried out under conditions of low stringency, moderate stringency or high stringency. As an example of nucleic acid hybridization, a polymer membrane containing immobilized denatured nucleic acids is first prehybridized for 30 minutes at 45□C in a solution consisting of 0.9 M NaCl, 50 mM NaH2PO4, pH 7.0, 5.0 mM Na2EDTA, 0.5% SDS, 10X Denhardt's, and 0.5 mg/ml polyriboadenylic acid. Approximately 2 X 107 cpm (specific activity 4-9 X 108 cpm/ug) of 32P end-labeled oligonucleotide probe are then added to the solution. After 12-16 hours of incubation, the membrane is washed for 30 minutes at room temperature in 1X SET (150 mM NaCl, 20 mM Tris hydrochloride, pH 7.8,1 mM Na2EDTA) containing 0.5% SDS, followed by a 30 minute wash in fresh 1X SET at Tm-10 C for the oligonucleotide probe. The membrane is then exposed to auto-radiographic film for detection of hybridization signals.

[0074] By varying the stringency of the hybridization conditions used to identify nucleic acids, such as cDNAs or genomic DNAs, which hybridize to the detectable probe, nucleic acids having different levels of homology to the probe can be identified and isolated. Stringency may be varied by conducting the hybridization at varying temperatures below the melting temperatures of the probes. The melting temperature, Tm, is the temperature (under defined ionic strenght and pH) at which 50% of the target sequence hybridizes to a perfectly complementary probe. Very stringent conditions are selected to be equal to or about 5 C lower than the Tm for a particular probe. The melting temperature of the probe may be calculated using the following formulas:

For probes between 14 and 70 nucleotides in length the melting temperature (Tm) is calculated using the formula: $Tm=81.5+16.6(\log [Na+])+0.41(\text{fraction G+C})-(600/N)$

where N is the length of the probe.

[0075] If the hybridization is carried out in a solution containing formamide, the melting temperature may be calculated using the equation: Tm=81.5+16.6(log [Na+])+0.41 (fraction G+C)-(0.63% formamide)-(600/N) where N is the length of the probe.

[0076] Prehybridization may be carried out in 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 g denatured fragmented salmon sperm DNA or 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 g denatured fragmented salmon sperm DNA, 50% formamide. The formulas for SSC and Denhardt's solutions are listed in Sambrook et al., supra.

[0077] Hybridization is conducted by adding the detectable probe to the prehybridization solutions listed above. Where the probe comprises double stranded DNA, it is denatured before addition to the hybridization solution. The filter is contacted with the hybridization solution for a sufficient period of time to allow the probe to hybridize to cDNAs or genomic DNAs containing sequences complementary thereto or homologous thereto. For probes over 200 nucleotides in length, the hybridization may be carried out at 15-25 C below the Tm. For shorter probes, such as oligonucleotide probes, the hybridization may be conducted at 5-10 C below the Tm. Typically, for hybridizations in 6X SSC, the hybridization is conducted at approximately 68 C. Usually, for hybridizations in 50% formamide containing solutions, the hybridization is conducted at approximately 42 C.

[0078] All of the foregoing hybridizations would be considered to be under conditions of high stringency.

[0079] Following hybridization, the filter is washed to remove any non-specifically bound detectable probe. The stringency used to wash the filters can also be varied depending on the nature of the nucleic acids being hybridized, the length of the nucleic acids being hybridized, the degree of complementarity, the nucleotide sequence composition (e.g., GC v. AT content), and the nucleic acid type (e.g., RNA v. DNA). Examples of progressively higher stringency condition washes are as follows: 2X SSC, 0.1% SDS at room temperature for 15 minutes (low stringency); 0.1X SSC, 0.5% SDS at room temperature for 30 minutes to 1 hour (moderate stringency); 0.1X SSC, 0.5% SDS for 15 to 30 minutes at between the hybridization temperature and 68°C (high stringency); and 0.15M NaCl for 15 minutes at 72°C (very high stringency). A final low stringency wash can be conducted in 0.1X SSC at room temperature. The examples above are merely illustrative of one set of conditions that can be used to wash filters. One of skill in the art would know that there are numerous recipes for different stringency washes. Some other examples are given below.

[0080] Nucleic acids which have hybridized to the probe are identified by autoradiography or other conventional techniques.

[0081] The above procedure may be modified to identify nucleic acids having decreasing levels of homology to the probe sequence. For example, to obtain nucleic acids of decreasing homology to the detectable probe, less stringent conditions may be used. For example, the hybridization temperature may be decreased in increments of 5°C from 68°C to 42°C in a hybridization buffer having a Na+ concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate" conditions above 50°C and "low" conditions below 50°C. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 55°C. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 45°C.

[0082] Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42°C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of homology to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 30% formamide. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 10% formamide.

[0083] For example, the preceding methods may be used to isolate nucleic acids having a sequence with at least about 97%, at least 95% or at least 90% homology to a nucleic acid sequence as set forth in SEQ ID No: 3, sequences substantially identical thereto, or fragments comprising at least about 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases thereof, and the sequences complementary to any of the foregoing sequences. Homology may be measured using an alignment algorithm. For example, the homologous polynucleotides may have a coding sequence which is a naturally occurring allelic variant of one of the coding sequences described herein. Such allelic variants may have a substitution, deletion or addition of one or more nucleotides when compared to the nucleic acid sequence as set forth in SEQ ID No: 3, or sequences complementary thereto.

[0084] Additionally, the above procedures may be used to isolate nucleic acids which encode polypeptides having at least about 99%, at least 95% or at least 90% homology to a polypeptide having a sequence as set forth in SEQ ID NO: 4, sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof as determined using a sequence alignment algorithm (e.g., such as the FASTA version 3.0t78 algorithm with the default parameters).

[0085] Another aspect of the invention is an isolated or purified polypeptide comprising a sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto. As discussed above, such polypeptides may be obtained

by inserting a nucleic acid encoding the polypeptide into a vector such that the coding sequence is operably linked to a sequence capable of driving the expression of the encoded polypeptide in a suitable host cell. For example, the expression vector may comprise a promoter, a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression.

[0086] Promoters suitable for expressing the polypeptide or fragment thereof in bacteria include the E. coli lac or trp promoters, the lacI promoter, the lacZ promoter, the T3 promoter, the T7 promoter, the gpt promoter, the lambda PR promoter, the lambda PL promoter, promoters from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), and the acid phosphatase promoter. Fungal promoters include the factor promoter. Eukaryotic promoters include the CMV immediate early promoter, the HSV thymidine kinase promoter, heat shock promoters, the early and late SV40 promoter, LTRs from retroviruses, and the mouse metallothionein-I promoter. Other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses may also be used.

[0087] Mammalian expression vectors may also comprise an origin of replication, any necessary ribosome binding sites, a polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. In some embodiments, DNA sequences derived from the SV40 splice and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

[0088] Vectors for expressing the polypeptide or fragment thereof in eukaryotic cells may also contain enhancers to increase expression levels. Enhancers are cis-acting elements of DNA, usually from about 10 to about 300 bp in length that act on a promoter to increase its transcription. Examples include the SV40 enhancer on the late side of the replication origin bp 100 to 270, the cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and the adenovirus enhancers.

[0089] In addition, the expression vectors typically contain one or more selectable marker genes to permit selection of host cells containing the vector. Such selectable markers include genes encoding dihydrofolate reductase or genes conferring neomycin resistance for eukaryotic cell culture, genes conferring tetracycline or ampicillin resistance in E. coli, and the S. cerevisiae TRP1 gene.

[0090] After the expression libraries have been generated one can include the additional step of "biopanning" such libraries prior to screening by cell sorting. The "biopanning" procedure refers to a process for identifying clones having a specified biological activity by screening for sequence homology in a library of clones prepared by (i) selectively isolating target DNA, from DNA derived from at least one microorganism, by use of at least one probe DNA comprising at least a portion of a DNA sequence encoding an biological having the specified biological activity; and (ii) optionally transforming a host with isolated target DNA to produce a library of clones which are screened for the specified biological activity.

[0091] The probe DNA used for selectively isolating the target DNA of interest from the DNA derived from at least one microorganism can be a full-length coding region sequence or a partial coding region sequence of DNA for an enzyme of known activity. The original DNA library can be preferably probed using mixtures of probes comprising at least a portion of the DNA sequence encoding an enzyme having the specified enzyme activity. These probes or probe libraries are preferably single-stranded and the microbial DNA which is probed has preferably been converted into single-stranded form. The probes that are particularly suitable are those derived from DNA encoding enzymes having an activity similar or identical to the specified enzyme activity which is to be screened.

[0092] The probe DNA should be at least about 10 bases and preferably at least 15 bases. In one embodiment, the entire coding region may be employed as a probe. Conditions for the hybridization in which target DNA is selectively isolated by the use of at least one DNA probe will be designed to provide a hybridization stringency of at least about 50% sequence identity, more particularly a stringency providing for a sequence identity of at least about 70%.

[0093] In nucleic acid hybridization reactions, the conditions used to achieve a particular level of stringency will vary, depending on the nature of the nucleic acids being hybridized. For example, the length, degree of complementarity, nucleotide sequence composition (e.g., GC v. AT content), and nucleic acid type (e.g., RNA v. DNA) of the hybridizing regions of the nucleic acids can be considered in selecting hybridization conditions. An additional consideration is whether one of the nucleic acids is immobilized, for example, on a filter.

[0094] An example of progressively higher stringency conditions is as follows: 2 x SSC/0.1% SDS at about room temperature (hybridization conditions); 0.2 x SSC/0.1% SDS at about room temperature (low stringency conditions); 0.2 x SSC/0.1% SDS at about 42°C (moderate stringency conditions); and 0.1 x SSC at about 68°C (high stringency conditions). Washing can be carried out using only one of these conditions, e.g., high stringency conditions, or each of the conditions can be used, e.g., for 10-15 minutes each, in the order listed above, repeating any or all of the steps listed. However, as mentioned above, optimal conditions will vary, depending on the particular hybridization reaction involved, and can be determined empirically.

[0095] Hybridization techniques for probing a microbial DNA library to isolate target DNA of potential interest are well known in the art and any of those which are described in the literature are suitable for use herein, particularly those which use a solid phase-bound, directly or indirectly bound, probe DNA for ease in separation from the remainder of the DNA derived from the microorganisms.

[0096] Preferably the probe DNA is "labeled" with one partner of a specific binding pair (i.e. a ligand) and the other

partner of the pair is bound to a solid matrix to provide ease of separation of target from its source. The ligand and specific binding partner can be selected from, in either orientation, the following: (1) an antigen or hapten and an antibody or specific binding fragment thereof; (2) biotin or iminobiotin and avidin or streptavidin; (3) a sugar and a lectin specific therefor; (4) an enzyme and an inhibitor therefor; (5) an apoenzyme and cofactor; (6) complementary homopolymeric oligonucleotides; and (7) a hormone and a receptor therefor. The solid phase is preferably selected from: (1) a glass or polymeric surface; (2) a packed column of polymeric beads; and (3) magnetic or paramagnetic particles.

[0097] Further, it is optional but desirable to perform an amplification of the target DNA that has been isolated. In this embodiment the target DNA is separated from the probe DNA after isolation. It is then amplified before being used to transform hosts. The double stranded DNA selected to include as at least a portion thereof a predetermined DNA sequence can be rendered single stranded, subjected to amplification and reannealed to provide amplified numbers of selected double stranded DNA. Numerous amplification methodologies are now well known in the art.

[0098] The selected DNA is then used for preparing a library for screening by transforming a suitable organism. Hosts, particularly those specifically identified herein as preferred, are transformed by artificial introduction of the vectors containing the target DNA by inoculation under conditions conducive for such transformation.

[0099] The resultant libraries of transformed clones are then screened for clones which display activity for the enzyme of interest.

[0100] Having prepared a multiplicity of clones from DNA selectively isolated from an organism, such clones are screened for a specific enzyme activity and to identify the clones having the specified enzyme characteristics.

[0101] The screening for enzyme activity may be effected on individual expression clones or may be initially effected on a mixture of expression clones to ascertain whether or not the mixture has one or more specified enzyme activities. If the mixture has a specified enzyme activity, then the individual clones may be rescreened utilizing a FACS machine for such enzyme activity or for a more specific activity. Alternatively, encapsulation techniques such as gel microdroplets, may be employed to localize multiple clones in one location to be screened on a FACS machine for positive expressing clones within the group of clones which can then be broken out into individual clones to be screened again on a FACS machine to identify positive individual clones. Thus, for example, if a clone mixture has hydrolase activity, then the individual clones may be recovered and screened utilizing a FACS machine to determine which of such clones has hydrolase activity. As used herein, "small insert library" means a gene library containing clones with random small size nucleic acid inserts of up to approximately 5000 base pairs. As used herein, "large insert library" means a gene library containing clones with random large size nucleic acid inserts of approximately 5000 up to several hundred thousand base pairs or greater.

[0102] As described with respect to one of the above aspects, the invention provides a process for enzyme activity screening of clones containing selected DNA derived from a microorganism which process includes: screening a library for specified enzyme activity, said library including a plurality of clones, said clones having been prepared by recovering from genomic DNA of a microorganism selected DNA, which DNA is selected by hybridization to at least one DNA sequence which is all or a portion of a DNA sequence encoding an enzyme having the specified activity; and transforming a host with the selected DNA to produce clones which are screened for the specified enzyme activity.

[0103] In one embodiment, a DNA library derived from a microorganism is subjected to a selection procedure to select therefrom DNA which hybridizes to one or more probe DNA sequences which is all or a portion of a DNA sequence encoding an enzyme having the specified enzyme activity by:

(a) rendering the double-stranded genomic DNA population into a single-stranded DNA population;
(b) contacting the single-stranded DNA population of (a) with the DNA probe bound to a ligand under conditions permissive of hybridization so as to produce a double-stranded complex of probe and members of the genomic DNA population which hybridize thereto;
(c) contacting the double-stranded complex of (b) with a solid phase specific binding partner for said ligand so as to produce a solid phase complex;
(d) separating the solid phase complex from the single-stranded DNA population of (b);
(e) releasing from the probe the members of the genomic population which had bound to the solid phase bound probe;
(f) forming double-stranded DNA from the members of the genomic population of (e);
(g) introducing the double-stranded DNA of (f) into a suitable host to form a library containing a plurality of clones containing the selected DNA; and
(h) screening the library for the specified enzyme activity.

[0104] In another aspect, the process includes a preselection to recover DNA including signal or secretion sequences. In this manner it is possible to select from the genomic DNA population by hybridization as hereinabove described only DNA which includes a signal or secretion sequence. The following paragraphs describe the protocol for this procedure, the nature and function of secretion signal sequences in general and a specific exemplary application of such sequences to an assay or selection process.

[0105] A particularly embodiment of this aspect further comprises, after (a) but before (b) above, the steps of:

(ai) contacting the single-stranded DNA population of (a) with a ligand-bound oligonucleotide probe that is complementary to a secretion signal sequence unique to a given class of proteins under conditions permissive of hybridization to form a double-stranded complex;
(aii) contacting the double-stranded complex of (ai) with a solid phase specific binding partner for said ligand so as to produce a solid phase complex;
(aiii) separating the solid phase complex from the single-stranded DNA population of (a);
(aiv) releasing the members of the genomic population which had bound to said solid phase bound probe; and
(av) separating the solid phase bound probe from the members of the genomic population which had bound thereto.

[0106] The DNA which has been selected and isolated to include a signal sequence is then subjected to the selection procedure hereinabove described to select and isolate therefrom DNA which binds to one or more probe DNA sequences derived from DNA encoding an enzyme(s) having the specified enzyme activity.

[0107] This procedure is described and exemplified in U.S. Serial No. 08/692,002, filed August 2, 1996.

[0108] In vivo biopanning may be performed utilizing a FACS-based machine. Complex gene libraries are constructed with vectors which contain elements which stabilize transcribed RNA. For example, the inclusion of sequences which result in secondary structures such as hairpins which are designed to flank the transcribed regions of the RNA would serve to enhance their stability, thus increasing their half life within the cell. The probe molecules used in the biopanning process consist of oligonucleotides labeled with reporter molecules that only fluoresce upon binding of the probe to a target molecule. These probes are introduced into the recombinant cells from the library using one of several transformation methods. The probe molecules bind to the transcribed target mRNA resulting in DNA/RNA heteroduplex molecules. Binding of the probe to a target will yield a fluorescent signal which is detected and sorted by the FACS machine during the screening process.

[0109] In some embodiments, the nucleic acid encoding the polypeptide of SEQ ID No: 4, sequences substantially identical thereto, or fragments comprising at least about 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is assembled in appropriate phase with a leader sequence capable of directing secretion of the translated polypeptide or fragment thereof. Optionally, the nucleic acid can encode a fusion polypeptide in which the polypeptide of SEQ ID No: 4, sequences substantially identical thereto, or fragments comprising at least 5, 10, 15,20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is fused to heterologous peptides or polypeptides, such as N-terminal identification peptides which impart desired characteristics, such as increased stability or simplified purification.

[0110] The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA sequence is ligated to the desired position in the vector following digestion of the insert and the vector with appropriate restriction endonucleases. Alternatively, blunt ends in both the insert and the vector may be ligated. A variety of cloning techniques are disclosed in Ausubel et al. Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook et al., Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989. Such procedures and others are deemed to be within the scope of those skilled in the art.

[0111] The vector may be, for example, in the form of a plasmid, a viral particle, or a phage. Other vectors include chromosomal, nonchromosomal and synthetic DNA sequences, derivatives of SV40; bacterial plasmids, phage DNA, baculovirus, yeast plasmids, vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies. A variety of cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989).

[0112] Particular bacterial vectors which may be used include the commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017), pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden), GEM1 (Promega Biotec, Madison, WI, USA) pQE70, pQE60, pQE-9 (Qiagen), pD10, psiX174 pBluescript II KS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene), ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia), pKK232-8 and pCM7. Particular eukaryotic vectors include pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). However, any other vector may be used as long as it is replicable and viable in the host cell.

[0113] The host cell may be any of the host cells familiar to those skilled in the art, including prokaryotic cells, eukaryotic cells, mammalian cells, insect cells, or plant cells. As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Bacillus subtilis, Salmonella typhimurium and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus, fungal cells, such as yeast, insect cells such as Drosophila S2 and Spodoptera Sf9, animal cells such as CHO, COS or Bowes melanoma, and adenoviruses. The selection of an appropriate host is within the abilities of those skilled in the art.

[0114] The vector may be introduced into the host cells using any of a variety of techniques, including transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer. Particular methods include calcium

phosphate transfection, DEAE-Dextran mediated transfection, lipofection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

[0115] Where appropriate, the engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the invention. Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter may be induced by appropriate means (e.g., temperature shift or chemical induction) and the cells may be cultured for an additional period to allow them to produce the desired polypeptide or fragment thereof.

[0116] Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract is retained for further purification. Microbial cells employed for expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known to those skilled in the art. The expressed polypeptide or fragment thereof can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the polypeptide. If desired, high performance liquid chromatography (HPLC) can be employed for final purification steps.

[0117] Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts (described by Gluzman, Cell, 23: 175, 1981), and other cell lines capable of expressing proteins from a compatible vector, such as the C127, 3T3, CHO, HeLa and BHK cell lines.

[0118] The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Depending upon the host employed in a recombinant production procedure, the polypeptides produced by host cells containing the vector may be glycosylated or may be non-glycosylated. Polypeptides of the invention may or may not also include an initial methionine amino acid residue.

[0119] Alternatively, the polypeptide of SEQ ID No 4, sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be synthetically produced by conventional peptide synthesizers. In other embodiments, fragments or portions of the polypeptides may be employed for producing the corresponding full-length polypeptide by peptide synthesis; therefore, the fragments may be employed as intermediates for producing the full-length polypeptides.

[0120] Cell-free translation systems can also be employed to produce the polypeptide of SEQ ID No 4, sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof using mRNAs transcribed from a DNA construct comprising a promoter operably linked to a nucleic acid encoding the polypeptide or fragment thereof. In some embodiments, the DNA construct may be linearized prior to conducting an in vitro transcription reaction. The transcribed mRNA is then incubated with an appropriate cell-free translation extract, such as a rabbit reticulocyte extract, to produce the desired polypeptide or fragment thereof.

[0121] The invention also relates to variants of the polypeptide of SEQ ID No 4, sequences substantially identical thereto. The term "variant" includes derivatives or analogs of these polypeptides. In particular, the variants may differ in amino acid sequence from the polypeptide of SEQ ID NO: 4, and sequences substantially identical thereto, by one or more substitutions, additions, deletions, fusions and truncations, which may be present in any combination.

[0122] The variants may be naturally occurring or created in vitro. In particular, such variants may be created using genetic engineering techniques such as site directed mutagenesis, random chemical mutagenesis, Exonuclease III deletion procedures, and standard cloning techniques. Alternatively, such variants, fragments, analogs, or derivatives may be created using chemical synthesis or modification procedures.

[0123] Other methods of making variants are also familiar to those skilled in the art. These include procedures in which nucleic acid sequences obtained from natural isolates are modified to generate nucleic acids which encode polypeptides having characteristics which enhance their value in industrial or laboratory applications. In such procedures, a large number of variant sequences having one or more nucleotide differences with respect to the sequence obtained from the natural isolate are generated and characterized. Typically, these nucleotide differences result in amino acid changes with respect to the polypeptides encoded by the nucleic acids from the natural isolates.

[0124] For example, variants may be created using error prone PCR. In error prone PCR, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Error prone PCR is described in Leung, D.W., et al., Technique, 1:11-15, 1989) and Caldwell, R. C. & Joyce G.F., PCR Methods Applic., 2:28-33, 1992. Briefly, in such procedures, nucleic acids to be mutagenized are mixed with PCR primers, reaction buffer, MgCl2, MnCl2, Taq polymerase and an appropriate concentration of dNTPs for achieving a high rate of point mutation along the entire length of the PCR product. For example, the reaction may be performed using 20 fmoles of nucleic acid to be mutagenized, 30pmole of each PCR primer, a reaction buffer comprising 50mM KCl, 10mM Tris HCl (pH 8.3) and 0.01% gelatin, 7mM MgCl2, 0.5mM MnCl2, 5 units of Taq polymerase, 0.2mM dGTP, 0.2mM dATP, 1mM dCTP, and 1mM dTTP. PCR may be performed for 30

cycles of 94°C for 1 min, 45° C for 1 min, and 72°C for 1 min. However, it will be appreciated that these parameters may be varied as appropriate. The mutagenized nucleic acids are cloned into an appropriate vector and the activities of the polypeptides encoded by the mutagenized nucleic acids is evaluated.

[0125] Variants may also be created using oligonucleotide directed mutagenesis to generate site-specific mutations in any cloned DNA of interest. Oligonucleotide mutagenesis is described in Reidhaar-Olson, J.F. & Sauer, R.T., et al., Science, 241:53-57, 1988. Briefly, in such procedures a plurality of double stranded oligonucleotides bearing one or more mutations to be introduced into the cloned DNA are synthesized and inserted into the cloned DNA to be mutagenized. Clones containing the mutagenized DNA are recovered and the activities of the polypeptides they encode are assessed.

[0126] Another method for generating variants is assembly PCR. Assembly PCR involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions occur in parallel in the same vial, with the products of one reaction priming the products of another reaction. Assembly PCR is described in pending U.S. Patent Application Serial No. 08/677,112 filed July 9, 1996, entitled, Method of "DNA Shuffling with Polynucleotides Produced by Blocking or interrupting a Synthesis or Amplification Process".

[0127] Still another method of generating variants is sexual PCR mutagenesis. In sexual PCR mutagenesis, forced homologous recombination occurs between DNA molecules of different but highly related DNA sequence in vitro, as a result of random fragmentation of the DNA molecule based on sequence homology, followed by fixation of the crossover by primer extension in a PCR reaction. Sexual PCR mutagenesis is described in Stemmer, W.P., PNAS, USA, 91: 10747-10751, 1994. Briefly, in such procedures a plurality of nucleic acids to be recombined are digested with DNAse to generate fragments having an average size of 50-200 nucleotides. Fragments of the desired average size are purified and resuspended in a PCR mixture. PCR is conducted under conditions which facilitate recombination between the nucleic acid fragments. For example, PCR may be performed by resuspending the purified fragments at a concentration of 10-30ng/ul in a solution of 0.2mM of each dNTP, 2.2mM MgCl2, 50mM KCL, 10mM Tris HCl, pH 9.0, and 0.1% Triton X-100. 2.5 units of Taq polymerase per 100□l of reaction mixture is added and PCR is performed using the following regime: 94 °C for 60 seconds, 94 °C for 30 seconds, 50-55° C for 30 seconds, 72 °C for 30 seconds (30-45 times) and 72° C for 5 minutes. However, it will be appreciated that these parameters may be varied as appropriate. In some embodiments, oligonucleotides may be included in the PCR reactions. In other embodiments, the Klenow fragment of DNA polymerase I may be used in a first set of PCR reactions and Taq polymerase may be used in a subsequent set of PCR reactions. Recombinant sequences are isolated and the activities of the polypeptides they encode are assessed.

[0128] Variants may also be created by in vivo mutagenesis. In some embodiments, random mutations in a sequence of interest are generated by propagating the sequence of interest in a bacterial strain, such as an E. coli strain, which carries mutations in one or more of the DNA repair pathways. Such "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in one of these strains will eventually generate random mutations within the DNA. Mutator strains suitable for use for in vivo mutagenesis are described in PCT Publication No. WO 91/16427, published October 31, 1991, entitled "Methods for Phenotype Creation from Multiple Gene Populations".

[0129] Variants may also be generated using cassette mutagenesis. In cassette mutagenesis a small region of a double stranded DNA molecule is replaced with a synthetic oligonucleotide "cassette" that differs from the native sequence. The oligonucleotide often contains completely and/or partially randomized native sequence.

[0130] Recursive ensemble mutagenesis may also be used to generate variants. Recursive ensemble mutagenesis is an algorithm for protein engineering (protein mutagenesis) developed to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. Recursive ensemble mutagenesis is described in Arkin, A.P. and Youvan, D.C., PNAS, USA, 89:7811-7815, 1992.

[0131] In some embodiments, variants are created using exponential ensemble mutagenesis. Exponential ensemble mutagenesis is a process for generating combinatorial libraries with a high percentage of unique and functional mutants, wherein small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Exponential ensemble mutagenesis is described in Delegrave, S. and Youvan, D.C., Biotechnology Research, 11:1548-1552, 1993. Random and site-directed mutagenesis are described in Arnold, F.H., Current Opinion in Biotechnology, 4:450-455, 1993.

In some embodiments, the variants are created using shuffling procedures wherein portions of a plurality of nucleic acids which encode distinct polypeptides are fused together to create chimeric nucleic acid sequences which encode chimeric polypeptides as described in pending U.S. Patent Application Serial No. 08/677,112 filed July 9, 1996, entitled, "Method of DNA Shuffling with Polynucleotides Produced by Blocking or interrupting a Synthesis or Amplification Process", and pending U.S. Patent Application Serial No. 08/651,568 filed May 22, 1996, entitled, "Combinatorial Enzyme Development".

[0132] The variants of the polypeptide of SEQ ID No: 4 may be variants in which one or more of the amino acid residues of the polypeptide of SEQ ID No: 4 are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code.

[0133] Conservative substitutions are those that substitute a given amino acid in a polypeptide by another amino acid of like characteristics. Typically seen as conservative substitutions are the following replacements: replacements of an aliphatic amino acid such as Ala, Val, Leu and Ile with another aliphatic amino acid; replacement of a Ser with a Thr or vice versa; replacement of an acidic residue such as Asp and Glu with another acidic residue; replacement of a residue bearing an amide group, such as Asn and Gln, with another residue bearing an amide group; exchange of a basic residue such as Lys and Arg with another basic residue; and replacement of an aromatic residue such as Phe, Tyr with another aromatic residue.

[0134] Other variants are those in which one or more of the amino acid residues of the polypeptide of SEQ ID No: 4 includes a substituent group.

[0135] Still other variants are those in which the polypeptide is associated with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol).

[0136] Additional variants are those in which additional amino acids are fused to the polypeptide, such as a leader sequence, a secretory sequence, a proprotein sequence or a sequence which facilitates purification, enrichment, or stabilization of the polypeptide.

[0137] In some embodiments, the fragments, derivatives and analogs retain the same biological function or activity as the polypeptide of SEQ ID No: 4, and sequences substantially identical thereto. In other embodiments, the fragment, derivative, or analog includes a proprotein, such that the fragment, derivative, or analog can be activated by cleavage of the proprotein portion to produce an active polypeptide.

[0138] Another aspect of the invention is polypeptides which have at least 90%, at least about 95%, or more than about 95% homology to the polypeptide of SEQ ID No: 4. Homology may be determined using any of the programs described above which aligns the polypeptides or fragments being compared and determines the extent of amino acid identity or similarity between them. It will be appreciated that amino acid "homology" includes conservative amino acid substitutions such as those described above.

[0139] The polypeptides or fragments having homology to the polypeptide of SEQ ID No: 4 may be obtained by isolating the nucleic acids encoding them using the techniques described above.

[0140] Alternatively, the homologous polypeptides or fragments may be obtained through biochemical enrichment or purification procedures. The sequence of potentially homologous polypeptides or fragments may be determined by proteolytic digestion, gel electrophoresis and/or microsequencing. The sequence of the prospective homologous polypeptide or fragment can be compared to the polypeptide of SEQ ID No: 4 using any of the programs described herein.

[0141] An assay may identify fragments or variants of SEQ ID No: 4, or sequences substantially identical thereto, which retain the enzymatic function of the polypeptide of SEQ ID No: 4, and sequences substantially identical thereto. For example the fragments or variants of the polypeptides, may be used to catalyze biochemical reactions, which indicate that said fragment or variant retains the enzymatic activity of the polypeptide in SEQ ID No: 4.

[0142] The assay for determining if fragments of variants retain the enzymatic activity of the polypeptide of SEQ ID No: 4, and sequences substantially identical thereto includes the steps of; contacting the polypeptide fragment or variant with a substrate molecule under conditions which allow the polypeptide fragment or variant to function, and detecting either a decrease in the level of substrate or an increase in the level of the specific reaction product of the reaction between the polypeptide and substrate.

The polypeptide of SEQ ID No: 4, sequences substantially identical thereto or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof may be used in a variety of applications. For example, the polypeptides or fragments thereof may be used to catalyze biochemical reactions. In accordance with one aspect of the invention, there is provided a process for utilizing a polypeptide of SEQ ID No: 4, and sequences substantially identical thereto or polynucleotides encoding such polypeptides for hydrolyzing haloalkanes. In such procedures, a substance containing a haloalkane compound is contacted with the polypeptide of SEQ ID No: 4, and sequences substantially identical thereto under conditions which facilitate the hydrolysis of the compound.

[0143] The polypeptide of SEQ ID No: 4, sequences substantially identical thereto or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof, may also be used to generate antibodies which bind specifically to the enzyme polypeptides or fragments. The resulting antibodies may be used in immunoaffinity chromatography procedures to isolate or purify the polypeptide or to determine whether the polypeptide is present in a biological sample. In such procedures, a protein preparation, such as an extract, or a biological sample is contacted with an antibody capable of specifically binding to the polypeptide of SEQ ID No: 4, sequences substantially identical thereto, or fragments of the foregoing sequences.

In immunoaffinity procedures, the antibody is attached to a solid support, such as a bead or other column matrix. The protein preparation is placed in contact with the antibody under conditions in which the antibody specifically binds to the polypeptide of SEQ ID No: 4, sequences substantially identical thereto, or fragment thereof. After a wash to remove non-specifically bound proteins, the specifically bound polypeptides are eluted.

The ability of proteins in a biological sample to bind to the antibody may be determined using any of a variety of procedures familiar to those skilled in the art. For example, binding may be determined by labeling the antibody with a detectable

label such as a fluorescent agent, an enzymatic label, or a radioisotope. Alternatively, binding of the antibody to the sample may be detected using a secondary antibody having such a detectable label thereon. Particular assays include ELISA assays, sandwich assays, radioimmunoassays, and Western Blots.

[0144] Polyclonal antibodies generated against the polypeptide of SEQ ID No: 4, and sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal, for example, a nonhuman. The antibody so obtained will then bind the polypeptide itself. In this manner, even a sequence encoding only a fragment of the polypeptide can be used to generate antibodies which may bind to the whole native polypeptide. Such antibodies can then be used to isolate the polypeptide from cells expressing that polypeptide.

[0145] For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kohler and Milstein, Nature, 256:495-497, 1975) the trioma technique, the human B-cell hybridoma technique (Kozbor et al., Immunology Today 4:72, 1983), and the EBV-hybridoma technique (Cole, et al., 1985, in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

[0146] Techniques described for the production of single chain antibodies (U.S. Patent No. 4,946,778) can be adapted to produce single chain antibodies to the polypeptide of SEQ ID No: 4, or fragments thereof. Alternatively, transgenic mice may be used to express humanized antibodies to these polypeptides or fragments.

[0147] Antibodies generated against a polypeptide of SEQ ID No: 4, sequences substantially identical thereto, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof may be used in screening for similar polypeptides from other organisms and samples. In such techniques, polypeptides from the organism are contacted with the antibody and those polypeptides which specifically bind the antibody are detected. Any of the procedures described above may be used to detect antibody binding. One such screening assay is described in "Methods for Measuring Cellulase Activities", Methods in Enzymology, Vol 160, pp. 87-116.

[0148] As used herein the term "nucleic acid sequence as set forth in SEQ ID No: 3" encompasses a nucleic acid sequence as set forth in SEQ ID No: 3, a sequence substantially identical to one of the foregoing sequences, fragments of any one or more of the foregoing sequences, nucleotide sequences homologous to SEQ ID No: 3 or homologous to fragments of SEQ ID No: 3, and sequences complementary to all of the preceding sequences. The fragments include portions of SEQ ID No: 3 comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive nucleotides of SEQ ID No: 3, and sequences substantially identical thereto. Homologous sequences and fragments of SEQ ID No: 3, and sequences substantially identical thereto, refer to a sequence having at least 99%, 98%, 97%, 96%, 95% or 90% homology to these sequences. Homology may be determined using any of the computer programs and parameters described herein, including FASTA version 3.0t7 with the default parameters. Homologous sequences also include RNA sequences in which uridines replace the thymines in the nucleic acid sequences as set forth in SEQ ID No: 3. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. It will be appreciated that the nucleic acid sequences of the invention can be represented in the traditional single character format (See the inside back cover of Stryer, Lubert. Biochemistry, 3rd edition. W. H Freeman & Co., New York.) or in any other format which records the identity of the nucleotides in a sequence.

[0149] As used herein the term "a polypeptide sequence as set forth in SEQ ID No: 4" encompasses a polypeptide sequence as set forth in SEQ ID No. : 4, sequences substantially identical thereto, which are encoded by a sequence as set forth in SEQ ID No: 3, polypeptide sequences homologous to the polypeptides of SEQ ID No: 4, and sequences substantially identical thereto, or fragments of any of the preceding sequences. Homologous polypeptide sequences refer to a polypeptide sequence having at least 99%, 98%, 97%, 96%, 95% or 90%, homology to one of the polypeptide sequences of the invention. Homology may be determined using any of the computer programs and parameters described herein, including FASTA version 3.0t78 with the default parameters or with any modified parameters. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. The polypeptide fragments comprise at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of the polypeptides of SEQ ID No: 4, and sequences substantially identical thereto. It will be appreciated that the polypeptides of the invention can be represented in the traditional single character format or three letter format (See the inside back cover of Starrier, Lubert. Biochemistry, 3rd edition. W. H Freeman & Co., New York.) or in any other format which relates the identity of the polypeptides in a sequence.

[0150] It will be appreciated by those skilled in the art that a nucleic acid sequence and a polypeptide sequence of the invention can be stored, recorded, and manipulated on any medium which can be read and accessed by a computer. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the nucleic acid sequences as set forth in SEQ ID No: 3, and sequences substantially identical thereto, one or more of the polypeptide sequences as set forth in SEQ ID No: 4, and sequences substantially identical thereto. Disclosed is also a computer readable medium having recorded thereon

at least 2, 5, 10, 15, or 20 nucleic acid sequences as set forth in SEQ ID No: 3, and sequences substantially identical thereto.

[0151] A computer readable medium may have recorded thereon one or more of the nucleic acid sequences as set forth in SEQ ID No: 3, and sequences substantially identical thereto. It may have recorded thereon one or more of the polypeptide sequences as set forth in SEQ ID No: 4, and sequences substantially identical thereto.

It may have recorded thereon at least 2, 5, 10, 15, or 20 of the sequences as set forth above.

[0152] Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM) as well as other types of other media known to those skilled in the art.

Systems (e.g., internet based systems), particularly computer systems, may store and manipulate the sequence information described herein. One example of a computer system 100 is illustrated in block diagram form. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze a nucleotide sequence of a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4. The computer system 100 typically includes a processor for processing, accessing and manipulating the sequence data. The processor 105 can be any well-known type of central processing unit, such as, for example, the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq, AMD or International Business Machines.

Typically the computer system 100 is a general purpose system that comprises the processor 105 and one or more internal data storage components 110 for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

The computer system 100 may include a processor 105 connected to a bus which is connected to a main memory 115 (preferably implemented as RAM) and one or more internal data storage devices 110, such as a hard drive and/or other computer readable media having data recorded thereon. The computer system 100 may further include one or more data retrieving device 118 for reading the data stored on the internal data storage devices 110.

The data retrieving device 118 may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, or a modem capable of connection to a remote data storage system (e.g., via the internet) etc. In some embodiments, the internal data storage device 110 is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system 100 may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system 100 includes a display 120 which is used to display output to a computer user. It should also be noted that the computer system 100 can be linked to other computer systems 125a-c in a network or wide area network to provide centralized access to the computer system 100.

Software for accessing and processing the nucleotide sequences of a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, (such as search tools, compare tools, and modeling tools etc.) may reside in main memory 115 during execution.

The computer system 100 may further comprise a sequence comparison algorithm for comparing a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, stored on a computer readable medium to a reference nucleotide or polypeptide sequence(s) stored on a computer readable medium. A "sequence comparison algorithm" refers to one or more programs which are implemented (locally or remotely) on the computer system 100 to compare a nucleotide sequence with other nucleotide sequences and/or compounds stored within a data storage means. For example, the sequence comparison algorithm may compare the nucleotide sequences of a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, stored on a computer readable medium to reference sequences stored on a computer readable medium to identify homologies or structural motifs. Various sequence comparison programs identified elsewhere in this patent specification are particularly contemplated. Protein and/or nucleic acid sequence homologies may be evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, Proc. Natl. Acad. Sci. USA 85(8):2444-2448, 1988; Altschul et al., J. Mol. Biol. 215 (3):403-410, 1990; Thompson et al., Nucleic Acids Res. 22(2):4673-4680, 1994; Higgins et al., Methods Enzymol. 266: 383-402, 1996; Altschul et al., J. Mol. Biol. 215(3):403-410, 1990; Altschul et al., Nature Genetics 3:266-272, 1993).

Homology or identity is often measured using sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, WI 53705). Such software matches similar sequences by assigning degrees of homology to various deletions, substitutions

and other modifications. The terms "homology" and "identity" in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same when compared and aligned for maximum correspondence over a comparison window or designated region as measured using any number of sequence comparison algorithms or by manual alignment and visual inspection. For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters. A "comparison window", as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 20 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequence for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482, 1981, by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol 48:443, 1970, by the search for similarity method of person & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444, 1988, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection. Other algorithms for determining homology or identity include, for example, in addition to a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information), ALIGN, AMAS (Analysis of Multiply Aligned Sequences), AMPS (Protein Multiple Sequence Alignment), ASSET (Aligned Segment Statistical Evaluation Tool), BANDS, BESTSCOR, BIOSCAN (Biological Sequence Comparative Analysis Node), BLIMPS (BLocks IMProved Searcher), FASTA, Intervals & Points, BMB, CLUSTAL V, CLUSTAL W, CONSENSUS, LCONSENSUS, WCONSENSUS, Smith-Waterman algorithm, DARWIN, Las Vegas algorithm, FNAT (Forced Nucleotide Alignment Tool), Framealign, Framesearch, DYNAMIC, FILTER, FSAP (Fristensky Sequence Analysis Package), GAP (Global Alignment Program), GENAL, GIBBS, GenQuest, ISSC (Sensitive Sequence Comparison), LALIGN (Local Sequence Alignment), LCP (Local Content Program), MACAW (Multiple Alignment Construction & Analysis Workbench), MAP (Multiple Alignment Program), MBLKP, MBLKN, PIMA (Pattern-Induced Multisequence Alignment), SAGA (Sequence Alignment by Genetic Algorithm) and WHAT-IF. Such alignment programs can also be used to screen genome databases to identify polynucleotide sequences having substantially identical sequences. A number of genome databases are available, for example, a substantial portion of the human genome is available as part of the Human Genome Sequencing Project (J. Roach, http://weber.u.Washington.edu/~roach/human_ genome_ progress 2.html) (Gibbs, 1995). At least twenty-one other genomes have already been sequenced, including, for example, M. genitalium (Fraser et al., 1995), M. jannaschii (Bult et al., 1996), H. influenzae (Fleischmann et al., 1995), E. coli (Blattner et al., 1997), and yeast (S. cerevisiae) (Mewes et al., 1997), and D. melanogaster (Adams et al., 2000). Significant progress has also been made in sequencing the genomes of model organism, such as mouse, C. elegans, and Arabadopsis sp. Several databases containing genomic information annotated with some functional information are maintained by different organization, and are accessible via the internet, for example, http://wwwtigr.org/tdb; http://www.genetics.wisc.edu; http://genome-www.stanford.edu/~ball; http://hiv-web.lanl.gov; http://www.ncbi.nlm.nih.gov; http://www.ebi.ac.uk; http://Pasteur.fr/other/biology; and http:// www.genome.wi.mit.edu. One example of a useful algorithm is BLAST and BLAST 2.0 algorithms, which are described in Altschul et al., Nuc. Acids Res. 25:3389-3402, 1977, and Altschul et al., J. Mol. Biol. 215:403-410, 1990, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al., supra). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectations (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, Proc. Natl. Acad. Sci. USA 89:10915, 1989) alignments (B) of 50, expectation (E) of 10, M=5, N= -4, and a comparison of both strands. The BLAST algorithm also performs a statistical analysis of the similarity

between two sequences (see, e.g., Karlin & Altschul, Proc. Natl. Acad. Sci. USA 90:5873, 1993). One measure of similarity provided by BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a references sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

In one embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;
(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;
(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;
(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and
(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet et al., Science 256:1443-1445, 1992; Henikoff and Henikoff, Proteins 17:49-61, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978, Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure, Washington: National Biomedical Research Foundation). BLAST programs are accessible through the U.S. National Library of Medicine, e.g., at www.ncbi.nlm.nih.gov.

[0153] The parameters used with the above algorithms may be adapted depending on the sequence length and degree of homology studied. In some embodiments, the parameters may be the default parameters used by the algorithms in the absence of instructions from the user.

[0154] The database of sequences can be a private database stored within the computer system 100, or a public database such as GENBANK that is available through the Internet.

[0155] The process 200 begins at a start state 201 and then moves to a state 202 wherein the new sequence to be compared is stored to a memory in a computer system 100. As discussed above, the memory could be any type of memory, including RAM or an internal storage device.

The process 200 then moves to a state 204 wherein a database of sequences is opened for analysis and comparison. The process 200 then moves to a state 206 wherein the first sequence stored in the database is read into a memory on the computer. A comparison is then performed at a state 210 to determine if the first sequence is the same as the second sequence. It is important to note that this step is not limited to performing an exact comparison between the new sequence and the first sequence in the database. Well-known methods are known to those of skill in the art for comparing two nucleotide or protein sequences, even if they are not identical. For example, gaps can be introduced into one sequence in order to raise the homology level between the two tested sequences. The parameters that control whether gaps or other features are introduced into a sequence during comparison are normally entered by the user of the computer system.

Once a comparison of the two sequences has been performed at the state 210, a determination is made at a decision state 210 whether the two sequences are the same. Of course, the term "same" is not limited to sequences that are absolutely identical. Sequences that are within the homology parameters entered by the user will be marked as "same" in the process 200.

If a determination is made that the two sequences are the same, the process 200 moves to a state 214 wherein the name of the sequence from the database is displayed to the user. This state notifies the user that the sequence with the displayed name fulfills the homology constraints that were entered. Once the name of the stored sequence is displayed to the user, the process 200 moves to a decision state 218 wherein a determination is made whether more sequences exist in the database. If no more sequences exist in the database, then the process 200 terminates at an end state 220. However, if more sequences do exist in the database, then the process 200 moves to a state 224 wherein a pointer is moved to the next sequence in the database so that it can be compared to the new sequence. In this manner, the new sequence is aligned and compared with every sequence in the database.

It should be noted that if a determination had been made at the decision state 212 that the sequences were not homologous, then the process 200 would move immediately to the decision state 218 in order to determine if any other sequences were available in the database for comparison.

A computer system may comprise a processor, a data storage device having stored thereon a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, a data storage device having retrievably stored thereon reference nucleotide sequences or polypeptide sequences to be compared to a nucleic acid sequence or a polypeptide sequence of the invention, and a sequence comparer for conducting the comparison. The sequence comparer may indicate a homology level between the sequences compared or identify structural motifs in the above described nucleic acid code of SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, or it may identify structural motifs in sequences which are compared to these nucleic acid codes and polypeptide codes. In some embodiments, the data storage device may have stored thereon the sequences of at least 2, 5, 10, 15, 20, 25, 30 or 40 or more of the nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or the polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto.

A method may determine the level of homology between a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, and a reference nucleotide sequence. The method includes reading the nucleic acid code or the polypeptide code and the reference nucleotide or polypeptide sequence through the use of a computer program which determines homology levels and determining homology between the nucleic acid code or polypeptide code and the reference nucleotide or polypeptide sequence with the computer program. The computer program may be any of a number of computer programs for determining homology levels, including those specifically enumerated herein, (e.g., BLAST2N with the default parameters or with any modified parameters). The method may be implemented using the computer systems described above. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30 or 40 or more of the above described nucleic acid sequence as set forth in SEQ ID No: 3 or the polypeptide sequences as set forth in SEQ ID No: 4 through use of the computer program and determining homology between the nucleic acid codes or polypeptide codes and reference nucleotide sequences or polypeptide sequences.

The process 250 begins at a start state 252 and then moves to a state 254 wherein a first sequence to be compared is stored to a memory. The second sequence to be compared is then stored to a memory at a state 256. The process 250 then moves to a state 260 wherein the first character in the first sequence is read and then to a state 262 wherein the first character of the second sequence is read. It should be understood that if the sequence is a nucleotide sequence, then the character would normally be either A, T, C, G or U. If the sequence is a protein sequence, then it is preferably in the single letter amino acid code so that the first and sequence sequences can be easily compared.

A determination is then made at a decision state 264 whether the two characters are the same. If they are the same, then the process 250 moves to a state 268 wherein the next characters in the first and second sequences are read. A determination is then made whether the next characters are the same. If they are, then the process 250 continues this loop until two characters are not the same. If a determination is made that the next two characters are not the same, the process 250 moves to a decision state 274 to determine whether there are any more characters either sequence to read.

If there are not any more characters to read, then the process 250 moves to a state 276 wherein the level of homology between the first and second sequences is displayed to the user. The level of homology is determined by calculating the proportion of characters between the sequences that were the same out of the total number of sequences in the first sequence. Thus, if every character in a first 100 nucleotide sequence aligned with a every character in a second sequence, the homology level would be 100%.

Alternatively, the computer program may be a computer program which compares the nucleotide sequences of a nucleic acid sequence as set forth in the invention, to one or more reference nucleotide sequences in order to determine whether the nucleic acid code of SEQ ID No: 3, and sequences substantially identical thereto, differs from a reference nucleic acid sequence at one or more positions. Optionally such a program records the length and identity of inserted, deleted or substituted nucleotides with respect to the sequence of either the reference polynucleotide or a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto. In one embodiment, the computer program may be a program which determines whether a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, contains a single nucleotide polymorphism (SNP) with respect to a reference nucleotide sequence.

A method may determine whether a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, differs at one or more nucleotides from a reference nucleotide sequence, comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through use of a computer program which identifies differences between nucleic acid sequences and identifying differences between the nucleic acid code and the reference nucleotide sequence with the computer program. The computer program may be a program which identifies single nucleotide polymorphisms. The method may be implemented by the computer systems described above and the method illustrated in Figure 3. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30, or 40 or more of the nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, and the reference

nucleotide sequences through the use of the computer program and identifying differences between the nucleic acid codes and the reference nucleotide sequences with the computer program.

The computer based system may further comprise an identifier for identifying features within a nucleic acid sequence as set forth in SEQ ID No: 3 or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto.

An "identifier" refers to one or more programs which identifies certain features within a nucleic acid sequence or a polypeptide sequence of the invention. In one embodiment, the identifier may comprise a program which identifies an open reading frame in a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto.

The process 300 begins at a start state 302 and then moves to a state 304 wherein a first sequence that is to be checked for features is stored to a memory 115 in the computer system 100. The process 300 then moves to a state 306 wherein a database of sequence features is opened. Such a database would include a list of each feature's attributes along with the name of the feature. For example, a feature name could be "Initiation Codon" and the attribute would be "ATG". Another example would be the feature name "TAATAA Box" and the feature attribute would be "TAATAA". An example of such a database is produced by the University of Wisconsin Genetics Computer Group (www.gcg.com). Alternatively, the features may be structural polypeptide motifs such as alpha helices, beta sheets, or functional polypeptide motifs such as enzymatic active sites, helix-turn-helix motifs or other motifs known to those skilled in the art.

Once the database of features is opened at the state 306, the process 300 moves to a state 308 wherein the first feature is read from the database. A comparison of the attribute of the first feature with the first sequence is then made at a state 310. A determination is then made at a decision state 316 whether the attribute of the feature was found in the first sequence. If the attribute was found, then the process 300 moves to a state 318 wherein the name of the found feature is displayed to the user.

[0156] The process 300 then moves to a decision state 320 wherein a determination is made whether move features exist in the database. If no more features do exist, then the process 300 terminates at an end state 324. However, if more features do exist in the database, then the process 300 reads the next sequence feature at a state 326 and loops back to the state 310 wherein the attribute of the next feature is compared against the first sequence.

It should be noted, that if the feature attribute is not found in the first sequence at the decision state 316, the process 300 moves directly to the decision state 320 in order to determine if any more features exist in the database.

A method may identify a feature within a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, comprising reading a nucleic acid sequence or a polypeptide sequence through the use of a computer program which identifies features therein and identifying features within the nucleic acid sequence or polypeptide sequence with the computer program. The computer program may comprise a computer program which identifies open reading frames. The method may be performed by reading a single sequence or at least 2, 5, 10, 15, 20, 25, 30, or 40 of the nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or the polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, through the use of the computer program and identifying features within the nucleic acid codes or polypeptide codes with the computer program.

In addition, a nucleic acid sequence or a polypeptide sequence of the invention may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto, may be stored as text in a word processing file, such as MicrosoftWORD or WORDPERFECT or as an ASCII file in a variety of database programs familiar to those of skill in the art, such as DB2, SYBASE, or ORACLE. In addition, many computer programs and databases may be used as sequence comparison algorithms, identifiers, or sources of reference nucleotide sequences or polypeptide sequences to be compared to a nucleic acid sequence as set forth in SEQ ID No: 3, and sequences substantially identical thereto, or a polypeptide sequence as set forth in SEQ ID No: 4, and sequences substantially identical thereto. The following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid sequences or the polypeptide sequences of the invention.

The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul et al, J. Mol. Biol. 215: 403, 1990), FASTA (Pearson and Lipman, Proc. Natl. Acad. Sci. USA, 85: 2444, 1988), FASTDB (Brutlag et al. Comp. App. Biosci. 6:237-245, 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius2.DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.), Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab Diversity Explorer (Molecular Simulations Inc.), Gene Explorer (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the MDL Available Chemicals Directory da-

tabase, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseqn database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

[0157] Each of these techniques is described in detail in the references mentioned. DNA can be mutagenized, or "evolved", utilizing any one or more of these techniques, and rescreened to identify more desirable clones. The invention will now be illustrated by the following working examples, which are in no way a limitation of the present invention.

[0158] The invention will now be described in greater detail by reference to the following non-limiting examples.

Examples

Example I

[0159] Phenylglycinonitrile was prepared using the Strecker reaction conditions as follows:

- 1 g 98% KCN in 4 ml H_2O in round bottomed flask
- 1.18 g NH_4Cl
- Mix at room temp until NH_4Cl has dissolved
- Add 2.12 g benzaldehyde in 4 ml methanol
- Stir for 2 h using a magnetic stirrer

The presence of the product, phenylglycinonitrile was detected by HPLC (7% isocratic MeOH in water; column: Supelcosil LC-18; 5 cm x 4.6 mm; 5 μm). After 2 hours, 70% conversion to product was obtained.

[0160] Various dilutions of the reaction mixture were made in 0.1 M sodium phosphate buffer (pH 7), ranging from no dilution to a fifty times dilution. Enzyme solutions were prepared by addition of buffer to lyophilized cell lysate preparations. These cell lysates contained nitrilase which had been overexpressed in a *Pseudomonas* host. Two enzyme preparations, example BD 1911 and example BD 1921, with final protein concentrations of 10.72 and 15.56 mg/ml of solution, were used. The enzyme solutions (20 μl) were added individually to each of the phenylglycinonitrile-containing dilutions (final volume 300 μl). The samples were incubated at room temperature. After 18 hours, the reactions were sampled and run on TLC (4:1:5 1-butanol: acetic acid: water). Phenylglycine was detected in the 10, 25 and 50 times dilutions, with a faint peak appearing in the 5 times dilution, for both enzymes. After 5 days, the reactions were sampled again and analyzed by HPLC. Nitrilase BD 1921 showed approximately 10% conversion to product in the 5 times dilution of substrate nitrilase BD 1911 converted < 5% of the substrate. Higher conversions were obtained for the lower dilutions of Strecker reaction mixture (up to 35% conversion in the 50 times dilution).

SEQUENCE LISTING

[0161]

<110> Diversa Corporation

<120> METHODS FOR PRODUCING ENANTIOMERICALLY PURE ALPHA-SUBSTITUTED CARBOXYLIC ACIDS

<130> DIVER1440WO

<140> Not yet known
<141> 2000-12-28

<150> 60/254,414
<151> 2000-12-07

<150> 60/173,609
<151> 1999-12-29

<160> 4

<170> PatentIn Ver. 2.1

<210> 1 .
<211> 1041
<212> DNA
<213> Unknown Organism

<220>
<223> Description of Unknown Organism: Obtained from an environmental sample

<220>
<221> CDS
<222> (1) .. (1041)

<400> 1

```
atg tcg gag ccc atg acg aag tat cgc ggc gcg gcg gtg cag gcc gcg   48
Met Ser Glu Pro Met Thr Lys Tyr Arg Gly Ala Ala Val Gln Ala Ala
  1               5                  10                  15

ccg gtg ttc ctc gat ctc gac cgc aca gtc gag aaa gcg atc ggc ctg   96
Pro Val Phe Leu Asp Leu Asp Arg Thr Val Glu Lys Ala Ile Gly Leu
             20                  25                  30

atc gag cag gcg gcc aag cag gac gtg cgc ctg atc gca ttc cca gag   144
Ile Glu Gln Ala Ala Lys Gln Asp Val Arg Leu Ile Ala Phe Pro Glu
         35                  40                  45

act tgg att ccc ggc tat ccc ttt tgg ata tgg ctg ggc gcg ccg gct   192
Thr Trp Ile Pro Gly Tyr Pro Phe Trp Ile Trp Leu Gly Ala Pro Ala
     50                  55                  60

tgg ggc atg cgc ttc gtc cag cgc tat ttc gag aat tcg ctc gtg cgc   240
Trp Gly Met Arg Phe Val Gln Arg Tyr Phe Glu Asn Ser Leu Val Arg
 65                  70                  75                  80

ggc agc aag cag tgg cag gcc ctg gcg gat gcg gcc cgc cgc cac ggc   288
Gly Ser Lys Gln Trp Gln Ala Leu Ala Asp Ala Ala Arg Arg His Gly
                 85                  90                  95
```

```
atg cat gtc gtg gcc ggc tat agc gag cgc gcg ggc ggc agc ctc tat   336
Met His Val Val Ala Gly Tyr Ser Glu Arg Ala Gly Gly Ser Leu Tyr
            100                 105                 110

atg ggc cag gcg atc ttc ggc ccc gat ggc gat ctg atc gcc gcg cgc   384
Met Gly Gln Ala Ile Phe Gly Pro Asp Gly Asp Leu Ile Ala Ala Arg
        115                 120                 125

cgc aag ctc aag cct acc cat gcg gag cgc acc gtg ttc ggc gag gga   432
Arg Lys Leu Lys Pro Thr His Ala Glu Arg Thr Val Phe Gly Glu Gly
    130                 135                 140

gac ggc agc cat ctc gcg gtg cac gat acc gcc atc ggg cgc ctc ggc   480
Asp Gly Ser His Leu Ala Val His Asp Thr Ala Ile Gly Arg Leu Gly
145                 150                 155                 160

gcg ctc tgt tgc tgg gag cac atc cag cca ttg tcg aaa tac gcc atg   528
Ala Leu Cys Cys Trp Glu His Ile Gln Pro Leu Ser Lys Tyr Ala Met
                165                 170                 175

tac gcc gcc gac gaa cag gtc cac gtc gcg tcg tgg ccg agc ttc agc   576
Tyr Ala Ala Asp Glu Gln Val His Val Ala Ser Trp Pro Ser Phe Ser
            180                 185                 190

ctc tat cgc ggc atg gcc tat gcg ctc gga ccg gag gtc aat acc gcc   624
Leu Tyr Arg Gly Met Ala Tyr Ala Leu Gly Pro Glu Val Asn Thr Ala
        195                 200                 205

gca agc cag atc tac gcg gtc gag ggc ggc tgc tac gtg ctg gcg tcg   672
Ala Ser Gln Ile Tyr Ala Val Glu Gly Gly Cys Tyr Val Leu Ala Ser
    210                 215                 220

tgc gcg acc gtt tcg ccg gag atg atc aag gta ttg gtg gat acg ccc   720
Cys Ala Thr Val Ser Pro Glu Met Ile Lys Val Leu Val Asp Thr Pro
225                 230                 235                 240

gac aag gag atg ttc ctc aag gcc ggc ggc ggt ttt gcc atg att ttc   768
Asp Lys Glu Met Phe Leu Lys Ala Gly Gly Gly Phe Ala Met Ile Phe
                245                 250                 255

ggg ccc gac ggc cgc gcc ctg gcc gag ccg ctc ccg gag acc gaa gag   816
Gly Pro Asp Gly Arg Ala Leu Ala Glu Pro Leu Pro Glu Thr Glu Glu
            260                 265                 270

gga ctg ctg gtc gcc gat atc gac ctc ggc atg atc gcg ttg gcc aag   864
Gly Leu Leu Val Ala Asp Ile Asp Leu Gly Met Ile Ala Leu Ala Lys
        275                 280                 285

gcg gcg gcc gat ccg gcg ggc cac tat tca cgg ccc gac gta acg cgg   912
Ala Ala Ala Asp Pro Ala Gly His Tyr Ser Arg Pro Asp Val Thr Arg
    290                 295                 300

ctg ctg ctg gat cga cgt ccg gcc caa cgc gtc gtc acg ctt gat gcc   960
Leu Leu Leu Asp Arg Arg Pro Ala Gln Arg Val Val Thr Leu Asp Ala
305                 310                 315                 320
```

```
gca ttc gaa ccg caa aac gag gac aag ggc gac gcg ccc gcg ctg cgc   1008
Ala Phe Glu Pro Gln Asn Glu Asp Lys Gly Asp Ala Pro Ala Leu Arg
                325                 330                 335

gtg gtg gcg gaa agc gcc gcc gcc gcg cag tag                      1041
Val Val Ala Glu Ser Ala Ala Ala Ala Gln
            340                 345
```

<210> 2
<211> 346
<212> PRT
<213> Unknown Organism

<223> Description of Unknown Organism: Obtained from an environmental sample

<400> 2

```
Met Ser Glu Pro Met Thr Lys Tyr Arg Gly Ala Ala Val Gln Ala Ala
  1               5                  10                  15
Pro Val Phe Leu Asp Leu Asp Arg Thr Val Glu Lys Ala Ile Gly Leu
             20                  25                  30
Ile Glu Gln Ala Ala Lys Gln Asp Val Arg Leu Ile Ala Phe Pro Glu
         35                  40                  45
Thr Trp Ile Pro Gly Tyr Pro Phe Trp Ile Trp Leu Gly Ala Pro Ala
     50                  55                  60
Trp Gly Met Arg Phe Val Gln Arg Tyr Phe Glu Asn Ser Leu Val Arg
 65                  70                  75                  80
Gly Ser Lys Gln Trp Gln Ala Leu Ala Asp Ala Ala Arg Arg His Gly
                 85                  90                  95
Met His Val Val Ala Gly Tyr Ser Glu Arg Ala Gly Gly Ser Leu Tyr
            100                 105                 110
Met Gly Gln Ala Ile Phe Gly Pro Asp Gly Asp Leu Ile Ala Ala Arg
        115                 120                 125
Arg Lys Leu Lys Pro Thr His Ala Glu Arg Thr Val Phe Gly Glu Gly
    130                 135                 140
Asp Gly Ser His Leu Ala Val His Asp Thr Ala Ile Gly Arg Leu Gly
145                 150                 155                 160
Ala Leu Cys Cys Trp Glu His Ile Gln Pro Leu Ser Lys Tyr Ala Met
                165                 170                 175
Tyr Ala Ala Asp Glu Gln Val His Val Ala Ser Trp Pro Ser Phe Ser
            180                 185                 190
Leu Tyr Arg Gly Met Ala Tyr Ala Leu Gly Pro Glu Val Asn Thr Ala
        195                 200                 205
Ala Ser Gln Ile Tyr Ala Val Glu Gly Gly Cys Tyr Val Leu Ala Ser
    210                 215                 220
Cys Ala Thr Val Ser Pro Glu Met Ile Lys Val Leu Val Asp Thr Pro
225                 230                 235                 240
Asp Lys Glu Met Phe Leu Lys Ala Gly Gly Gly Phe Ala Met Ile Phe
                245                 250                 255
Gly Pro Asp Gly Arg Ala Leu Ala Glu Pro Leu Pro Glu Thr Glu Glu
            260                 265                 270
Gly Leu Leu Val Ala Asp Ile Asp Leu Gly Met Ile Ala Leu Ala Lys
        275                 280                 285
Ala Ala Ala Asp Pro Ala Gly His Tyr Ser Arg Pro Asp Val Thr Arg
    290                 295                 300
Leu Leu Leu Asp Arg Arg Pro Ala Gln Arg Val Val Thr Leu Asp Ala
305                 310                 315                 320
Ala Phe Glu Pro Gln Asn Glu Asp Lys Gly Asp Ala Pro Ala Leu Arg
                325                 330                 335
```

Val Val Ala Glu Ser Ala Ala Ala Ala Gln
340 345

<210> 3
<211> 1014
<212> DNA
<213> Unknown Organism

<220>
<223> Description of Unknown Organism: Obtained from an environmental sample

<220>
<221> CDS
<222> (1) .. (1014)

<400> 3

```
atg aaa gaa gct atc aag gtc gcc tgc gtg caa gcc gcc ccg atc tac   48
Met Lys Glu Ala Ile Lys Val Ala Cys Val Gln Ala Ala Pro Ile Tyr
  1               5                  10                  15

atg gat ttg gag gcg acg gtg gac aaa acc att gag ttg atg gaa gaa   96
Met Asp Leu Glu Ala Thr Val Asp Lys Thr Ile Glu Leu Met Glu Glu
             20                  25                  30

gca gca cgt aat aat gct cgt ctg atc gcc ttt ccg gaa act tgg att   144
Ala Ala Arg Asn Asn Ala Arg Leu Ile Ala Phe Pro Glu Thr Trp Ile
         35                  40                  45

cca ggc tac cca tgg ttt ctt tgg ctt gac tca cca gca tgg gca atg   192
Pro Gly Tyr Pro Trp Phe Leu Trp Leu Asp Ser Pro Ala Trp Ala Met
     50                  55                  60

caa ttt gta cgc caa tac cat gag aac tca ttg gag ttg gat ggc cct   240
Gln Phe Val Arg Gln Tyr His Glu Asn Ser Leu Glu Leu Asp Gly Pro
 65                  70                  75                  80

caa gct aag cgc att tca gat gca gcc aag cgg ttg gga atc atg gtc   288
Gln Ala Lys Arg Ile Ser Asp Ala Ala Lys Arg Leu Gly Ile Met Val
                 85                  90                  95

acc ctg ggg atg agt gaa cgg gtc ggt ggc acc ctt tac atc agt cag   336
Thr Leu Gly Met Ser Glu Arg Val Gly Gly Thr Leu Tyr Ile Ser Gln
            100                 105                 110

tgg ttc ata ggc gat aat ggt gac acc att ggg gcc cgg cga aag ttg   384
Trp Phe Ile Gly Asp Asn Gly Asp Thr Ile Gly Ala Arg Arg Lys Leu
        115                 120                 125

aaa cct act ttt gtt gaa cgt act ttg ttc ggc gaa ggg gat ggt tca   432
Lys Pro Thr Phe Val Glu Arg Thr Leu Phe Gly Glu Gly Asp Gly Ser
    130                 135                 140

tcg cta gcg gtt ttc gag acg tct gtt gga agg ctg ggt ggc tta tgc   480
Ser Leu Ala Val Phe Glu Thr Ser Val Gly Arg Leu Gly Gly Leu Cys
145                 150                 155                 160
```

```
tgt tgg gag cac ctt caa ccg cta aca aaa tac gct ttg tat gca caa   528
Cys Trp Glu His Leu Gln Pro Leu Thr Lys Tyr Ala Leu Tyr Ala Gln
                165                 170                 175

aat gaa gag att cat tgt gcg gct tgg ccg agc ttt agc ctt tat cct   576
Asn Glu Glu Ile His Cys Ala Ala Trp Pro Ser Phe Ser Leu Tyr Pro
            180                 185                 190

aat gcg gcg aaa gcc ctg ggg cct gat gtc aat gta gcg gcc tct cga   624
Asn Ala Ala Lys Ala Leu Gly Pro Asp Val Asn Val Ala Ala Ser Arg
        195                 200                 205

atc tat gcc gtt gaa ggg caa tgc ttc gta cta gcg tcg tgt gcg ctc   672
Ile Tyr Ala Val Glu Gly Gln Cys Phe Val Leu Ala Ser Cys Ala Leu
    210                 215                 220

gtt tca caa tcc atg atc gat atg ctt tgt aca gat gac gaa aag cat   720
Val Ser Gln Ser Met Ile Asp Met Leu Cys Thr Asp Asp Glu Lys His
225                 230                 235                 240

gcg ttg ctt ctg gct ggt ggt gga cac tca cgt atc ata ggg cct gat   768
Ala Leu Leu Leu Ala Gly Gly Gly His Ser Arg Ile Ile Gly Pro Asp
                245                 250                 255

ggt ggt gac ttg gtc gcg cct ctt gcc gaa aat gaa gag ggt att ctc   816
Gly Gly Asp Leu Val Ala Pro Leu Ala Glu Asn Glu Glu Gly Ile Leu
            260                 265                 270

tac gca aac ctt gat cct gga gta cgc atc ctt gct aaa atg gcg gca   864
Tyr Ala Asn Leu Asp Pro Gly Val Arg Ile Leu Ala Lys Met Ala Ala
        275                 280                 285

gac cct gct ggt cat tat tcc cgt ccc gac att act cgc ttg cta ata   912
Asp Pro Ala Gly His Tyr Ser Arg Pro Asp Ile Thr Arg Leu Leu Ile
    290                 295                 300

gat cgc agc cct aaa tta ccg gta gtt gaa att gaa ggt gat ctt cgt   960
Asp Arg Ser Pro Lys Leu Pro Val Val Glu Ile Glu Gly Asp Leu Arg
305                 310                 315                 320

cct tac gct ttg ggt aaa gcg tct gag acg ggt gcg caa ctc gaa gaa  1008
Pro Tyr Ala Leu Gly Lys Ala Ser Glu Thr Gly Ala Gln Leu Glu Glu
                325                 330                 335

att tga                                                         1014
Ile
```

<210> 4
<211> 337
<212> PRT
<213> Unknown Organism

<223> Description of Unknown Organism: obtained from an environmental sample

<400> 4

```
Met Lys Glu Ala Ile Lys Val Ala Cys Val Gln Ala Ala Pro Ile Tyr
 1               5                  10                  15
```

```
Met Asp Leu Glu Ala Thr Val Asp Lys Thr Ile Glu Leu Met Glu Glu
                 20                  25                  30
Ala Ala Arg Asn Asn Ala Arg Leu Ile Ala Phe Pro Glu Thr Trp Ile
             35                  40                  45
Pro Gly Tyr Pro Trp Phe Leu Trp Leu Asp Ser Pro Ala Trp Ala Met
         50                  55                  60
Gln Phe Val Arg Gln Tyr His Glu Asn Ser Leu Glu Leu Asp Gly Pro
 65                  70                  75                  80
Gln Ala Lys Arg Ile Ser Asp Ala Ala Lys Arg Leu Gly Ile Met Val
                 85                  90                  95
Thr Leu Gly Met Ser Glu Arg Val Gly Gly Thr Leu Tyr Ile Ser Gln
            100                 105                 110
Trp Phe Ile Gly Asp Asn Gly Asp Thr Ile Gly Ala Arg Arg Lys Leu
        115                 120                 125
Lys Pro Thr Phe Val Glu Arg Thr Leu Phe Gly Glu Gly Asp Gly Ser
    130                 135                 140
Ser Leu Ala Val Phe Glu Thr Ser Val Gly Arg Leu Gly Gly Leu Cys
145                 150                 155                 160
Cys Trp Glu His Leu Gln Pro Leu Thr Lys Tyr Ala Leu Tyr Ala Gln
                165                 170                 175
Asn Glu Glu Ile His Cys Ala Ala Trp Pro Ser Phe Ser Leu Tyr Pro
            180                 185                 190
Asn Ala Ala Lys Ala Leu Gly Pro Asp Val Asn Val Ala Ala Ser Arg
        195                 200                 205
Ile Tyr Ala Val Glu Gly Gln Cys Phe Val Leu Ala Ser Cys Ala Leu
    210                 215                 220
Val Ser Gln Ser Met Ile Asp Met Leu Cys Thr Asp Asp Glu Lys His
225                 230                 235                 240
Ala Leu Leu Leu Ala Gly Gly Gly His Ser Arg Ile Ile Gly Pro Asp
                245                 250                 255
Gly Gly Asp Leu Val Ala Pro Leu Ala Glu Asn Glu Glu Gly Ile Leu
            260                 265                 270
Tyr Ala Asn Leu Asp Pro Gly Val Arg Ile Leu Ala Lys Met Ala Ala
        275                 280                 285
Asp Pro Ala Gly His Tyr Ser Arg Pro Asp Ile Thr Arg Leu Leu Ile
    290                 295                 300
Asp Arg Ser Pro Lys Leu Pro Val Val Glu Ile Glu Gly Asp Leu Arg
305                 310                 315                 320
Pro Tyr Ala Leu Gly Lys Ala Ser Glu Thr Gly Ala Gln Leu Glu Glu
                325                 330                 335
Ile
```

Claims

1. A substantially purified, isolated or recombinant polypeptide having an amino acid sequence

   (a) comprising a sequence as set forth in SEQ ID NO:4, or
   (b) comprising sequences having at least 90%, sequence identity to SEQ ID NO:4, wherein the polypeptide has nitrilase activity.

2. An isolated or recombinant nucleic acid comprising

   (a) a polynucleotide sequence encoding a polypeptide having (i) an amino acid sequence as set forth in SEQ ID NO:4, or (ii) amino acid sequences having at least 90% sequence identity to SEQ ID NO:4, wherein the polypeptide has nitrilase activity;
   (b) a polynucleotide sequence that hybridizes to the nucleic acid sequence as set forth in SEQ ID NO:3, wherein the polynucleotide encoding a polypeptide having nitrilase activity, and the hybridization conditions comprise high stringency conditions comprising 42°C in 50% formamide, 5X SSPE, 0.3% SDS, and 200 n/ml sheared

and denatured salmon sperm DNA;
(c) a polynucleotide sequence having at least 90% sequence identity to SEQ ID NO:3 over the coding region, wherein the polynucleotide encodes a polypeptide having nitrilase activity; or,
(d) a polynucleotide sequence complementary to (a), (b) or (c).

3. A vector comprising a nucleic acid as set forth in claim 2, wherein optionally the vector is an expression vector or a cloning vector, and optionally the vector comprises a cosmid, a fosmid, a bacterial artificial chromosome (BAC), a baculovirus, a phage, a plasmid, a phagemid, a viral DNA, a P1-based artificial chromosomes, a yeast plasmid or a yeast artificial chromosome.

4. A host cell comprising a polypeptide as set forth in claim 1, a nucleic acid as set forth in claim 2 or a vector as set forth in claim 3, wherein optionally the cell is a prokaryotic cell, a eukaryotic cell, a mammalian cell, a plant cell, a fungal cell, a yeast cell or an insect cell.

5. An isolated or recombinant antibody capable of specifically binding to a polypeptide as set forth in claim 1, or a polypeptide encoded by a nucleic acid as set forth in claim 2.

6. A method for hydrolyzing amino nitriles or cyanohydrin intermediates to produce an α-substituted carboxylic acid, said method comprising
contacting an aldehyde or ketone with a cyanide containing compound and an ammonia-containing compound or an ammonium salt or an amine, and
hydrolyzing the resulting amino nitrile or cyanohydrin intermediate with a nitrilase or a polypeptide having nitrilase activity, wherein the nitrilase is sufficiently active to perform the hydrolysis in the presence of the reaction components, under conditions and for a time sufficient to produce the carboxylic acid,
wherein the nitrilase or a polypeptide having nitrilase activity

(a) comprises a polypeptide according to claim 1; or
(b) is encoded by a nucleic acid according to claim 2.

7. A method for producing an alpha-substituted carboxylic acid, the method comprising

(a) providing an aldehyde or a ketone;
(b) providing a cyanide-containing compound;
(c) providing an ammonia-containing compound or a compound comprising an ammonium salt or an amine;
(d) providing a composition comprising a nitrilase or a polypeptide having nitrilase activity which comprises a polypeptide according to claim 1 or is encoded by a nucleic acid according to claim 2;
(e) contacting the aldehyde or ketone of step (a) with a cyanide-containing compound of step (b) and an ammonia-containing compound or a compound comprising an ammonium salt or an amine of step (c) such that an amino nitrile or a cyanohydrin intermediate is produced; and
(f) contacting the amino nitrile or cyanohydrin intermediate of step (e) with the composition of step (d) such that the nitrilase or polypeptide having nitrilase activity hydrolyzes the amino nitrile or cyanohydrin intermediate to produce an alpha-substituted carboxylic acid.

8. A method for producing an alpha-substituted carboxylic acid, the method comprising

(a) providing a composition comprising an amino nitrile or a cyanohydrin;
(b) providing a composition comprising a nitrilase or a polypeptide having nitrilase activity, which comprises a polypeptide according to claim 1 or is encoded by a nucleic acid according to claim 2; and
(c) contacting the amino nitrile or cyanohydrin of step (a) with the composition of step (b) such that the nitrilase or polypeptide having nitrilase activity hydrolyzes the amino nitrile or cyanohydrin intermediate to produce an alpha-substituted carboxylic acid.

9. A method for producing an alpha-amino acid, the method comprising

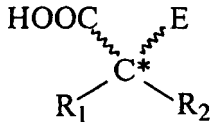
(a) providing an aldehyde or a ketone;
(b) providing a cyanide-containing compound and ammonia;
(c) providing a composition comprising a nitrilase or a polypeptide having nitrilase activity, which comprises a polypeptide according to claim 1 or is encoded by a nucleic acid according to claim 2;

(d) contacting the aldehyde or ketone of step (a) with the cyanide-containing compound and ammonia of step (b) such that an amino nitrile is produced; and
(e) contacting the amino nitrile of step (d) with the nitrilase or the polypeptide having nitrilase activity of step (c) such that the nitrilase or the polypeptide having nitrilase activity hydrolyzes the amino nitrile to produce an alpha-substituted amino acid.

10. A method for producing an alpha-substituted carboxylic acid, said method comprising
contacting an aldehyde or ketone with a cyanide-containing compound and an ammonia-containing compound, an ammonium salt or an amine, and
hydrolyzing the resulting amino nitrile or cyanohydrin intermediate with a nitrilase or a polypeptide having nitrilase activity, wherein the nitrilase hydrolyzes the reaction components to produce an alpha-substituted carboxylic acid and wherein the nitrilase or polypeptide having nitrilase activity comprises a polypeptide according to claim 1 or is encoded by a nucleic acid according to claim 2.

11. A method for producing an alpha-substituted carboxylic acid, the method comprising
providing a composition comprising a nitrilase or a polypeptide having nitrilase activity, which comprises a polypeptide according to claim 1 or is encoded by a nucleic acid according to claim 2; and
contacting reaction components with the composition such that the nitrilase or polypeptide having nitrilase activity hydrolyzes the reaction components to produce an alpha-substituted carboxylic acid,
wherein the reaction components are an aldehyde or ketone, a cyanide-containing compound, and an ammonia-containing compound, ammonia salt, or amine.

12. The method according to any one of claims 6 to 11, wherein said α-substituted carboxylic acid has the following structure:



wherein:

R_1 and R_2 are each independently -H, substituted or unsubstituted alkyl, alkenyl, alkynyl, aryl, heteroaryl, cycloalkyl, heterocyclic, wherein said substituents are lower alkyl, hydroxy, alkoxy, mercapto, cycloalkyl, heterocyclic, aryl, heteroaryl, aryloxy, or halogen or optionally R_1 and R_2 are linked to cooperate to form a functional cyclic moiety and E is $-N(R_x)_2$ or -OH, wherein each R_x is -H or lower alkyl.

13. The method according to claim 12, wherein said α-substituted carboxylic acid is an α-amino acid.

14. The method according to claim 12 or 13, wherein at least one of R_1 and R_2 is substituted or unsubstituted aryl.

15. The method according to claim 13 or 14, wherein said α-amino acid is D-phenylalanine, D-phenylglycine, or L-methylphenylglycine.

16. The method according to claim 13, wherein said α-amino acid bears a substituted or unsubstituted alkyl side chain.

17. The method according to claim 16, wherein said α-amino acid is L-tert-leucine, D-alanine, or D-hydroxynorleucine.

18. The method according to claim 12, wherein said α-substituted carboxylic acid is an α-hydroxy acid.

19. The method according to claim 18, wherein at least one of R_1 and R_2 is substituted or unsubstituted aryl.

20. The method according to claim 18, wherein said α-hydroxy acid is (S)-cyclohexylmandelic acid, mandelic acid or 2-chloro mandelic acid.

21. The method according to any one of claims 6 to 11, wherein the cyanide is a metal cyanide or a gaseous cyanide.

22. The method according to claim 21, wherein the cyanide is an alkali cyanide.

23. The method according to claim 21, wherein the metal cyanide is sodium cyanide.

24. The method according to any one of claims 6 to 11, wherein the ammonium salt has the formula $NH_2(R)_2^+X$, wherein each R is independently -H or lower alkyl, and X is a counter ion.

25. The method according to claim 24, wherein X is a halide.

26. The method according to claim 25, wherein the halide is Cl^-.

27. The method according to claim 26, wherein the ammonium salt is $NH_4^+Cl^-$.

28. The method of any one of claims 6 to 27, wherein the method produces an enantiomerically pure α-substituted carboxylic acid or an enantiomerically pure α-substituted amino acid.

29. The method of any one of claims 6 to 11, wherein the reaction takes place in a single reaction vessel.

30. A method of generating a variant of a nucleic acid encoding a polypeptide with an enzyme activity comprising the steps of:

(a) providing a template nucleic acid comprising a sequence as set forth in claim 2; and
(b) modifying, deleting or adding one or more nucleotides in the template sequence, or a combination thereof, to generate a variant of the template nucleic acid encoding a polypeptide with retained enzyme activity or a polypeptide including a proprotein, cleavage of which produces a polypeptide with enzyme activity,

wherein optionally the method further comprises expressing the variant nucleic acid to generate a variant polypeptide; and optionally the variants can be produced by a method comprising error-prone polymerase chain reaction (PCR), shuffling, oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, *in vivo* mutagenesis, cassette mutagenesis, recursive ensemble mutagenesis, exponential ensemble mutagenesis, site-specific mutagenesis, ligation reassembly, gene site saturation mutagenesis (GSSM), synthetic ligation reassembly (SLR) and any combination thereof.

31. A method of producing a recombinant polypeptide comprising the steps of: (a) providing a nucleic acid operably linked to a promoter, wherein the nucleic acid comprises a sequence as set forth in claim 2; and (b) expressing the nucleic acid of step (a) under conditions that allow expression of the polypeptide, thereby producing a recombinant polypeptide;
wherein optionally the method further comprises transforming a host cell with the nucleic acid of step (a) followed by expressing the nucleic acid of step (a), thereby producing a recombinant polypeptide in a transformed cell.

32. The method of claim 6, 7, 8 or claim 10, wherein the nitrilase or polypeptide having nitrilase activity stereoselectively hydrolyzes the resulting amino nitrile or cyanohydrin intermediate.

33. The method of claim 9, wherein the nitrilase or the polypeptide having nitrilase activity stereoselectively hydrolyzes the amino nitrile.

34. The method of claim 11, wherein the nitrilase or polypeptide having nitrilase activity stereoselectively hydrolyzes the reaction components.

Patentansprüche

1. Im Wesentlichen gereinigtes, isoliertes oder rekombinantes Polypeptid mit einer Aminosäuresequenz, die

(a) eine Sequenz wie in der SEQ ID NO: 4 dargelegt beinhaltet, oder
(b) Sequenzen beinhaltet, die wenigstens 90 % Sequenzidentität zur SEQ ID NO: 4 haben, wobei das Polypeptid Nitrilaseaktivität hat.

2. Isolierte oder rekombinante Nucleinsäure, die Folgendes beinhaltet:

(a) eine Polynucleotidsequenz, die ein Polypeptid mit (i) einer Aminosäuresequenz wie in SEQ ID NO: 4 dargelegt oder mit (ii) Aminosäuresequenzen mit wenigstens 90 % Sequenzidentität zur SEQ ID NO: 4 kodiert, wobei das Polypeptid Nitrilaseaktivität hat;
(b) eine Polynucleotidsequenz, die mit der Nucleinsäuresequenz wie in SEQ ID NO: 3 dargelegt hybridisiert, wobei das Polynucleotid ein Polypeptid mit Nitrilaseaktivität kodiert und die Hybridisierungsbedingungen hochstringente Bedingungen von 42°C in 50 % Formamid, 5X SSPE, 0,3 % SDS und 200 n/ml gescherte und denaturierte Lachsspermien-DNA beinhalten;
(c) eine Polynucleotidsequenz mit wenigstens 90 % Sequenzidentität zur SEQ ID NO: 3 über die Kodierungsregion, wobei das Polynucleotid ein Polypeptid mit Nitrilaseaktivität kodiert; oder
(d) eine Polynucleotidsequenz, die komplementär zu (a), (b) oder (c) ist.

3. Vektor, der eine Nucleinsäure nach Anspruch 2 beinhaltet, wobei der Vektor optional ein Expressionsvektor oder ein Klonierungsvektor ist und der Vektor optional ein Cosmid, ein Fosmid, ein bakterielles künstliches Chromosom (BAC), ein Baculovirus, ein Phage, ein Plasmid, ein Phagemid, eine Virus-DNA, ein P1-basiertes künstliches Chromosom, ein Hefe-Plasmid oder ein künstliches Hefechromosom ist.

4. Wirtszelle, die ein Polypeptid nach Anspruch 1, eine Nucleinsäure nach Anspruch 2 oder einen Vektor nach Anspruch 3 beinhaltet, wobei die Zelle optional eine prokaryotische Zelle, eine eukaryotische Zelle, eine Säugetierzelle, eine Pflanzenzelle, eine Pilzzelle, eine Hefezelle oder eine Insektenzelle ist.

5. Isolierter oder rekombinanter Antikörper, der sich spezifisch an ein Polypeptid nach Anspruch 1 oder ein Polypeptid binden kann, das durch eine Nucleinsäure nach Anspruch 2 kodiert ist.

6. Verfahren zum Hydrolysieren von Aminonitrilen oder Cyanohydrinintermediaten, um eine α-substituierte Carbonsäure zu produzieren, wobei das genannte Verfahren Folgendes beinhaltet:

Inkontaktbringen eines Aldehyds oder Ketons mit einer cyanidhaltigen Verbindung und einer ammoniakhaltigen Verbindung oder einem Ammoniumsalz oder einem Amin, und
Hydrolysieren des resultierenden Aminonitrils oder Cyanohydrinintermediats mit einer Nitrilase oder einem Polypeptid mit Nitrilaseaktivität, wobei die Nitrilase ausreichend aktiv ist, um die Hydrolyse in Anwesenheit der Reaktionskomponenten durchzuführen, unter Bedingungen und über einen Zeitraum, die/der ausreicht, um die Carbonsäure zu produzieren,
wobei die Nitrilase oder ein Polypeptid mit Nitrilaseaktivität

(a) ein Polypeptid nach Anspruch 1 beinhaltet; oder
(b) durch eine Nucleinsäure nach Anspruch 2 kodiert ist.

7. Verfahren zum Produzieren einer alpha-substituierten Carbonsäure, wobei das Verfahren Folgendes beinhaltet:

(a) Bereitstellen eines Aldehyds oder eines Ketons;
(b) Bereitstellen einer cyanidhaltigen Verbindung;
(c) Bereitstellen einer ammoniakhaltigen Verbindung oder einer Verbindung, die ein Ammoniumsalz oder ein Amin enthält;
(d) Bereitstellen einer Zusammensetzung, die eine Nitrilase oder ein Polypeptid mit einer Nitrilaseaktivität beinhaltet, die/das ein Polypeptid nach Anspruch 1 beinhaltet oder durch eine Nucleinsäure nach Anspruch 2 kodiert ist;
(e) Inkontaktbringen des Aldehyds oder Ketons aus Schritt (a) mit einer cyanidhaltigen Verbindung aus Schritt (b) und einer ammoniakhaltigen Verbindung oder einer ein Ammoniumsalz oder ein Amin enthaltenden Verbindung aus Schritt (c), so dass ein Aminonitril oder ein Cyanohydrinintermediat produziert wird; und
(f) Inkontaktbringen des Aminonitrils oder Cyanohydrinintermediats aus Schritt (e) mit der Zusammensetzung aus Schritt (d), so dass die Nitrilase oder das Polypeptid mit Nitrilaseaktivität das Aminonitril oder Cyanohydrinintermediat hydrolysiert, um eine alpha-substituierte Carbonsäure zu produzieren.

8. Verfahren zum Produzieren einer alpha-substituierten Carbonsäure, wobei das Verfahren Folgendes beinhaltet:

(a) Bereitstellen einer Zusammensetzung, die ein Aminonitril oder ein Cyanohydrin beinhaltet;

(b) Bereitstellen einer Zusammensetzung, die eine Nitrilase oder ein Polypeptid mit Nitrilaseaktivität beinhaltet, die/das ein Polypeptid nach Anspruch 1 beinhaltet oder durch eine Nucleinsäure nach Anspruch 2 kodiert ist; und
(c) Inkontaktbringen des Aminonitrils oder Cyanohydrins aus Schritt (a) mit der Zusammensetzung aus Schritt (b), so dass die Nitrilase oder das Polypeptid mit Nitrilaseaktivität das Aminonitril oder Cyanohydrinintermediat hydrolysiert, um eine alpha-substituierte Carbonsäure zu produzieren.

9. Verfahren zum Produzieren einer alpha-Aminosäure, wobei das Verfahren Folgendes beinhaltet:

(a) Bereitstellen eines Aldehyds oder eines Ketons;
(b) Bereitstellen einer cyanidhaltigen Verbindung und von Ammoniak;
(c) Bereitstellen einer Zusammensetzung, die eine Nitrilase oder ein Polypeptid mit Nitrilaseaktivität beinhaltet, die/das ein Polypeptid nach Anspruch 1 beinhaltet oder durch eine Nucleinsäure nach Anspruch 2 kodiert ist;
(d) Inkontaktbringen des Aldehyds oder Ketons aus Schritt (a) mit der cyanidhaltigen Verbindung und dem Ammoniak aus Schritt (b), so dass ein Aminonitril produziert wird; und
(e) Inkontaktbringen des Aminonitrils aus Schritt (d) mit der Nitrilase oder dem Polypeptid mit Nitrilaseaktivität aus Schritt (c), so dass die Nitrilase oder das Polypeptid mit Nitrilaseaktivität das Aminonitril hydrolysiert, um eine alpha-substituierte Aminosäure zu produzieren.

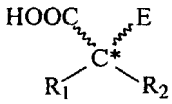
10. Verfahren zum Produzieren einer alpha-substituierten Carbonsäure, wobei das genannte Verfahren Folgendes beinhaltet:

Inkontaktbringen eines Aldehyds oder Ketons mit einer cyanidhaltigen Verbindung und einer ammoniakhaltigen Verbindung, einem Ammoniumsalz oder einem Amin, und
Hydrolysieren des resultierenden Aminonitrils oder Cyanohydrinintermediats mit einer Nitrilase oder einem Polypeptid mit Nitrilaseaktivität, wobei die Nitrilase die Reaktionskomponenten hydrolysiert, um eine alpha-substituierte Carbonsäure zu produzieren, und
wobei die Nitrilase oder das Polypeptid mit Nitrilaseaktivität ein Polypeptid nach Anspruch 1 beinhaltet oder durch eine Nucleinsäure nach Anspruch 2 kodiert ist.

11. Verfahren zum Produzieren einer alpha-substituierten Carbonsäure, wobei das Verfahren Folgendes beinhaltet:

Bereitstellen einer Zusammensetzung, die eine Nitrilase oder ein Polypeptid mit Nitrilaseaktivität beinhaltet, die/das ein Polypeptid nach Anspruch 1 beinhaltet oder durch eine Nucleinsäure nach Anspruch 2 kodiert ist; und
Inkontaktbringen der Reaktionskomponenten mit der Zusammensetzung, so dass die Nitrilase oder das Polypeptid mit Nitrilaseaktivität die Reaktionskomponenten hydrolysiert, um eine alpha-substituierte Carbonsäure zu produzieren,
wobei die Reaktionskomponenten ein Aldehyd oder Keton, eine cyanidhaltige Verbindung und eine ammoniakhaltige Verbindung, ein Ammoniaksalz oder Amin sind.

12. Verfahren nach einem der Ansprüche 6 bis 11, wobei die genannte α-substituierte Carbonsäure die folgende Struktur hat:



wobei:

R_1 und R_2 jeweils unabhängig -H, substituiertes oder unsubstituiertes Alkyl, Alkenyl, Alkynyl, Aryl, Heteroaryl, Cycloalkyl, Heterocyclus sind, wobei die genannten Substituenten niederes Alkyl, Hydroxy, Alkoxy, Mercapto, Cycloalkyl, Heterocyclus, Aryl, Heteroaryl, Aryloxy oder Halogen sind oder R_1 und R_2 optional zum Zusammenwirken verknüpft sind, um einen funktionellen cyclischen Anteil zu bilden, und E $-N(R_x)_2$ oder -OH ist, wobei jedes R_x -H oder niederes Alkyl ist.

13. Verfahren nach Anspruch 12, wobei die genannte α-substituierte Carbonsäure eine α-Aminosäure ist.

14. Verfahren nach Anspruch 12 oder 13, wobei R_1 und/oder R_2 substituiertes oder unsubstituiertes Aryl ist/sind.

15. Verfahren nach Anspruch 13 oder 14, wobei die genannte α-Aminosäure D-Phenylalanin, D-Phenylglycin oder L-Methylphenylglycin ist.

16. Verfahren nach Anspruch 13, wobei die genannte α-Aminosäure eine substituierte oder unsubstituierte Alkylseitenkette trägt.

17. Verfahren nach Anspruch 16, wobei die genannte α-Aminosäure L-tert-Leucin, D-Alanin oder D-Hydroxynorleucin ist.

18. Verfahren nach Anspruch 12, wobei die genannte α-substituierte Carbonsäure eine α-Hydroxysäure ist.

19. Verfahren nach Anspruch 18, wobei R_1 und/oder R_2 substituiertes oder unsubstituiertes Aryl ist/sind.

20. Verfahren nach Anspruch 18, wobei die genannte α-Hydroxysäure (S)-Cyclohexylmandelsäure, Mandelsäure oder 2-Chlormandelsäure ist.

21. Verfahren nach einem der Ansprüche 6 bis 11, wobei das Cyanid ein Metallcyanid oder ein gasförmiges Cyanid ist.

22. Verfahren nach Anspruch 21, wobei das Cyanid ein Alkalicyanid ist.

23. Verfahren nach Anspruch 21, wobei das Metallcyanid Natriumcyanid ist.

24. Verfahren nach einem der Ansprüche 6 bis 11, wobei das Ammoniumsalz die Formel $NH_2(R)_2^+X$ hat, wobei jedes R unabhängig -H oder niederes Alkyl ist und X ein Gegenion ist.

25. Verfahren nach Anspruch 24, wobei X ein Halogenid ist.

26. Verfahren nach Anspruch 25, wobei das Halogenid Cl^- ist.

27. Verfahren nach Anspruch 26, wobei das Ammoniumsalz $NH_4^+Cl^-$ ist.

28. Verfahren nach einem der Ansprüche 6 bis 27, wobei das Verfahren eine enantiomerenreine α-substituierte Carbonsäure oder eine enantiomerenreine α-substituierte Aminosäure hervorbringt.

29. Verfahren nach einem der Ansprüche 6 bis 11, wobei die Reaktion in einem einzelnen Reaktionsgefäß stattfindet.

30. Verfahren zum Erzeugen einer Variante einer Nucleinsäure, die ein Polypeptid mit einer Enzymaktivität kodiert, das die folgenden Schritte beinhaltet:

(a) Bereitstellen einer Matrizennucleinsäure, die eine Sequenz nach Anspruch 2 beinhaltet; und
(b) Modifizieren, Deletieren oder Addieren von einem oder mehreren Nucleotiden in der Matrizensequenz, oder eine Kombination davon, zum Erzeugen einer Variante der Matrizennucleinsäure, die ein Polypeptid mit beibehaltener Enzymaktivität oder ein Polypeptid kodiert, das ein Proprotein enthält, wobei eine Spaltung davon ein Polypeptid mit Enzymaktivität hervorbringt,

wobei das Verfahren optional ferner das Exprimieren der Nucleinsäurevariante beinhaltet, um eine Polypeptidvariante zu erzeugen;
und die Varianten optional mit einem Verfahren produziert werden können, das fehleranfällige Polymerasekettenreaktion (PCR), Shuffling, Oligonucleotidgerichtete Mutagenese, Assemblierungs-PCR, Sexual-PCR-Mutagenese, *In-vivo*-Mutagenese, Kassettenmutagenese, Recursive-Ensemble-Mutagenese, Exponential-Ensemble-Mutagenese, ortsspezifische Mutagenese, Ligationsreassemblierung, GSSM (Gene Site Saturation Mutagenesis), SLR (Synthetic Ligation Reassembly) und eine beliebige Kombination davon beinhaltet.

31. Verfahren zum Produzieren eines rekombinanten Polypeptids, das die folgenden Schritte beinhaltet:

(a) Bereitstellen einer Nucleinsäure, die mit einem Promotor funktionsfähig verknüpft ist, wobei die Nucleinsäure eine Sequenz nach Anspruch 2 beinhaltet; und

(b) Exprimieren der Nucleinsäure aus Schritt (a) unter Bedingungen, die eine Expression des Polypeptids ermöglichen, so dass ein rekombinantes Polypeptid produziert wird;

wobei das Verfahren optional ferner das Transformieren einer Wirtszelle mit der Nucleinsäure aus Schritt (a) beinhaltet, gefolgt vom Exprimieren der Nucleinsäure aus Schritt (a), wodurch ein rekombinantes Polypeptid in einer transformierten Zelle produziert wird.

32. Verfahren nach Anspruch 6, 7, 8 oder 10, wobei die Nitrilase oder das Polypeptid mit Nitrilaseaktivität das resultierende Aminonitril oder Cyanohydrinintermediat stereoselektiv hydrolysiert.

33. Verfahren nach Anspruch 9, wobei die Nitrilase oder das Polypeptid mit Nitrilaseaktivität das Aminonitril stereoselektiv hydrolysiert.

34. Verfahren nach Anspruch 11, wobei die Nitrilase oder das Polypeptid mit Nitrilaseaktivität die Reaktionskomponenten stereoselektiv hydrolysiert.

Revendications

1. Polypeptide isolé ou recombinant sensiblement purifié ayant une séquence d'acides aminés

(a) comprenant une séquence selon la SEQ ID n°4, ou
(b) comprenant des séquences ayant au moins 90% d'identité de séquence avec la SEQ ID n°4, le polypeptide ayant une activité nitrilase.

2. Acide nucléique isolé ou recombinant comprenant

(a) une séquence de polynucléotides codant un polypeptide ayant (i) une séquence d'acides aminés selon la SEQ ID n°4, ou (ii) des séquences d'acides aminés ayant au moins 90% d'identité de séquence avec la SEQ ID n°4, le polypeptide ayant une activité nitrilase ;
(b) une séquence de polynucléotides qui s'hybride à la séquence d'acide nucléique selon la SEQ ID n°3, le polynucléotide codant un polypeptide ayant une activité nitrilase, et les conditions d'hybridation comprenant des conditions de stringence élevée avec 42 °C dans 50% de formamide, 5 X SSPE, 0,3% de SDS, et 200n/ml d'ADN de sperme de saumon cisaillé et dénaturé ;
(c) une séquence de polynucléotides ayant au moins 90% d'identité de séquence avec la SEQ ID n°3 sur la région de codage, le polynucléotide codant un polypeptide ayant une activité nitrilase ; ou
(d) une séquence de polynucléotides complémentaire à (a), (b) ou (c).

3. Vecteur comprenant un acide nucléique selon la revendication 2, le vecteur étant éventuellement un vecteur d'expression ou un vecteur de clonage, et éventuellement le vecteur comprenant un cosmide, un fosmide, un chromosome bactérien artificiel (BAC), un baculovirus, un phage, un plasmide, un phagémide, un ADN viral, un chromosome artificiel dérivé du phage P1, un plasmide de levure ou un chromosome artificiel de levure.

4. Cellule hôte comprenant un polypeptide selon la revendication 1, un acide nucléique selon la revendication 2 ou un vecteur selon la revendication 3, la cellule étant éventuellement une cellule procaryote, une cellule eucaryote, une cellule de mammifère, une cellule végétale, une cellule fongique, une cellule de levure ou une cellule d'insecte.

5. Anticorps isolé ou recombinant capable de se lier spécifiquement à un polypeptide selon la revendication 1, ou un polypeptide codé par un acide nucléique selon la revendication 2.

6. Procédé d'hydrolyse d'aminonitriles ou d'intermédiaires de cyanohydrine afin de produire un acide carboxylique alpha-substitué, le procédé comprenant
la mise en contact d'un aldéhyde ou d'une cétone avec un composé à teneur en cyanure et un composé en teneur en ammoniac ou un sel d'ammonium ou une amine, et
l'hydrolyse de l'aminonitrile ou de l'intermédiaire de cyanohydrine résultant avec une nitrilase ou un polypeptide ayant une activité nitrilase, dans lequel la nitrilase est suffisamment active pour exécuter l'hydrolyse en présence des composants de réaction, dans des conditions et pendant une durée suffisante pour produire l'acide carboxylique, dans lequel la nitrilase ou un polypeptide ayant une activité nitrilase

(a) comprend un polypeptide selon la revendication 1 ; ou
(b) est codé par un acide nucléique selon la revendication 2.

7. Procédé de production d'un acide carboxylique alpha-substitué, le procédé comprenant

(a) la fourniture d'un aldéhyde ou d'une cétone ;
(b) la fourniture d'un composé à teneur en cyanure ;
(c) la fourniture d'un composé à teneur en ammoniac ou d'un composé comprenant un sel d'ammonium ou une amine ;
(d) la fourniture d'une composition comprenant une nitrilase ou un polypeptide ayant une activité nitrilase, laquelle comprend un polpypeptide selon la revendication 1 ou est codée par un acide nucléique selon la revendication 2 ;
(e) la mise en contact de l'aldéhyde ou de la cétone de l'étape (a) avec un composé à teneur en cyanure de l'étape (b) et un composé en teneur en ammoniac ou un composé comprenant un sel d'ammonium ou une amine selon l'étape (c) de façon à produire un aminonitrile ou un intermédiaire de cyanohydrine ; et
(f) la mise en contact de l'aminonitrile ou de l'intermédiaire de cyanohydrine selon l'étape (e) avec la composition de l'étape (d) de telle sorte que la nitrilase ou le polypeptide ayant une activité nitrilase hydrolyse l'aminonitrile ou l'intermédiaire de cyanohydrine afin de produire un acide carboxylique alpha-substitué.

8. Procédé de production d'un acide carboxylique alpha-substitué, le procédé comprenant

(a) la fourniture d'une composition comprenant un aminonitrile ou une cyanohydrine ;
(b) la fourniture d'une composition comprenant une nitrilase ou un polypeptide ayant une activité nitrilase, laquelle comprend un polpypeptide selon la revendication 1 ou est codée par un acide nucléique selon la revendication 2 ; et
(c) la mise en contact de l'aminonitrile ou de la cyanohydrine de l'étape (a) avec la composition de l'étape (b) de telle sorte que la nitrilase ou le polypeptide à activité nitrilase hydrolyse l'aminonitrile ou l'intermédiaire de cyanohydrine afin de produire un acide carboxylique alpha-substitué.

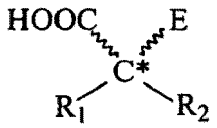
9. Procédé de production d'un alpha-aminoacide, le procédé comprenant

(a) la fourniture d'un aldéhyde ou d'une cétone ;
(b) la fourniture d'un composé à teneur en cyanure et ammoniac ;
(c) la fourniture d'une composition comprenant une nitrilase ou un polypeptide ayant une activité nitrilase, laquelle comprend un polpypeptide selon la revendication 1 ou est codée par un acide nucléique selon la revendication 2 ;
(d) la mise en contact de l'aldéhyde ou de la cétone de l'étape (a) avec un composé à teneur en cyanure et ammoniac de l'étape (b) de façon à produire un aminonitrile ; et
(e) la mise en contact de l'aminonitrile de l'étape (d) avec la nitrilase ou le polypeptide a activité nitrilase selon l'étape (c) de telle sorte que la nitrilase ou le polypeptide à activité nitrilase hydrolyse l'aminonitrile afin de produire un aminoacide alpha-substitué.

10. Procédé de production d'un acide carboxylique alpha-substitué, ledit procédé comprenant la mise en contact d'un aldéhyde ou d'une cétone avec un composé à teneur en cyanure et un composé en teneur en ammoniac, un sel d'ammonium ou une amine, et l'hydrolyse de l'aminonitrile ou de l'intermédiaire de cyanohydrine résultant avec une nitrilase ou un polypeptide ayant une activité nitrilase, la nitrilase hydrolysant les composants de réaction afin de produire un acide carboxylique alpha-substitué et
dans lequel la nitrilase ou le polypeptide à activité nitrilase comprend un polypeptide selon la revendication 1 ou est codé par un acide nucléique selon la revendication 2.

11. Procédé de production d'un acide carboxylique alpha-substitué, le procédé comprenant
la fourniture d'une composition comprenant une nitrilase ou un polypeptide ayant une activité nitrilase, laquelle comprend un polpypeptide selon la revendication 1 ou est codée par un acide nucléique selon la revendication 2 ; et
la mise en contact de composants de réaction avec la composition de telle sorte que la nitrilase ou le polypeptide à activité nitrilase hydrolyse les composants de réaction afin de produire un acide carboxylique alpha-substitué,
dans lequel les composants de réaction sont un aldéhyde ou une cétone, un composé à teneur en cyanure, et un composant à teneur en ammoniac, un sel d'ammonium ou une amine.

12. Procédé selon l'une quelconque des revendications 6 à 11, dans lequel ledit acide carboxylique alpha-substitué a la structure suivante :



où :

R_1 et R_2 représentent chacun indépendamment l'hydrogène, un groupe alkyle, alcényle, alcynile, aryle, hétéroaryle, cycloalkyle, hétérocyclique, substitué ou non substitué, dans lequel lesdits substituants sont un groupe alkyle, hydroxy, alkoxy, mercapto, cycloalkyle, hétérocyclique, aryle, hétéroaryle, aryloxy ou halogène inférieur ou éventuellement R_1 et R_2 sont liés pour coopérer et former un groupe caractéristique cyclique fonctionnel et E est $-N(R_x)_2$ ou -OH, dans lequel chaque R_x est de l'hydrogène ou un alkyle inférieur.

13. Procédé selon la revendication 12, dans lequel ledit acide carboxylique alpha-substitué est un alpha-aminoacide.

14. Procédé selon la revendication 12 ou 13, dans lequel au moins l'un de R_1 et R_2 est un aryle substitué ou non substitué.

15. Procédé selon la revendication 13 ou 14, dans lequel ledit alpha-aminoacide est D-phénylalanine, D-phénylglycine ou L-méthylphénylglycine.

16. Procédé selon la revendication 13, dans lequel ledit alpha-aminoacide porte une chaîne latérale à alkyle substitué ou non substitué.

17. Procédé selon la revendication 16, dans lequel ledit alpha-aminoacide est L-tert leucine, D-alanine ou D-hydroxynorleucine.

18. Procédé selon la revendication 12, dans lequel ledit acide carboxylique alpha-substitué est un acide alpha-hydroxylé.

19. Procédé selon la revendication 18, dans lequel au moins l'un de R_1 et R_2 est un groupe aryle substitué ou non substitué.

20. Procédé selon la revendication 18, dans lequel ledit acide alpha-hydroxylé est un acide (S)-cyclohexylmandélique, un acide mandélique ou un acide 2-chloro mandélique.

21. Procédé selon l'une quelconque des revendications 6 à 11, dans lequel le cyanure est un cyanure métallique ou un cyanure gazeux.

22. Procédé selon la revendication 21, dans lequel le cyanure est un cyanure alcalin.

23. Procédé selon la revendication 21, dans lequel le cyanure métallique est un cyanure de sodium.

24. Procédé selon l'une quelconque des revendications 6 à 11, dans lequel le sel d'ammonium a la formule $NH_2(R)_2{}^+X$, où chaque R est indépendamment de l'hydrogène ou un alkyle inférieur, et X est un contre-ion.

25. Procédé selon la revendication 24, dans lequel X est un halogénure.

26. Procédé selon la revendication 25, dans lequel l'halogénure est Cl^-.

27. Procédé selon la revendication 26, dans lequel le sel d'ammonium est $NH_4{}^+Cl^-$.

28. Procédé selon l'une quelconque des revendications 6 à 27, le procédé produisant un acide carboxylique alpha-substitué énantiomériquement pur ou un acide aminé alpha-substitué énantiomériquement pur.

29. Procédé selon l'une quelconque des revendications 6 à 11, dans lequel la réaction se déroule dans une même cuve de réaction.

30. Procédé de génération d'un variant d'un acide nucléique codant un polypeptide ayant une activité enzymatique comprenant les étapes consistant à :

(a) fournir une matrice d'acide nucléique comprenant une séquence selon la revendication 2 ; et
(b) modifier, supprimer ou ajouter un ou plusieurs nucléotides dans la matrice de séquence, ou une combinaison de ces opérations, pour générer une variante de la matrice d'acide nucléique codant un polypeptide à activité enzymatique conservée ou un polypeptide comportant une proprotéine, dont le clivage produit un polypeptide à activité enzymatique, éventuellement le procédé comprenant en outre l'expression du variant d'acide nucléique pour générer un variant de polypeptide ;

et éventuellement les variants pouvant être produits par un procédé comprenant la réaction en chaîne par polymérase (PCR) sujette à l'erreur, le brassage, la mutagenèse dirigée par oligonucléotide, la PCR d'assemblage, la mutagenèse par PCR sexuelle, la mutagenèse *in vivo,* la mutagenèse par cassettes, la mutagenèse par ensemble récursif, la mutagenèse par ensemble exponentiel, la mutagenèse spécifique au site, le réassemblage par ligature, la mutagenèse à saturation de site de gènes (GSSM), le réassemblage synthétique par ligature (SLR), et n'importe quelle combinaison de ceux-ci.

31. Procédé de production d'un polypeptide recombinant comprenant les étapes consistant à : (a) fournir un acide nucléique lié opérationnellement à un promoteur, dans lequel l'acide nucléique comprend une séquence selon la revendication 2 ; et (b) exprimer l'acide nucléique de l'étape (a) dans des conditions qui permettent l'expression du polypeptide, produisant ainsi un polypeptide recombinant ;
le procédé comprenant éventuellement en outre la transformation d'une cellule hôte avec l'acide nucléique de l'étape (a) puis l'expression de l'acide nucléique de l'étape (a), produisant ainsi un polypeptide recombinant dans une cellule transformée.

32. Procédé selon la revendication 6, 7, 8 ou 10, dans lequel la nitrilase ou le polypeptide à activité nitrilase hydrolyse stéréosélectivement l'aminonitrile ou l'intermédiaire de cyanohydrine résultant.

33. Procédé selon la revendication 9, dans lequel la nitrilase ou le polypeptide à activité nitrilase hydrolyse stéréosélectivement l'aminonitrile.

34. Procédé selon la revendication 11, dans lequel la nitrilase ou le polypeptide à activité nitrilase hydrolyse stéréosélectivement les composants de réaction.

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO 8607386 A **[0005]**
- EP 0610048 A **[0006]**
- EP 0449648 A **[0007]**
- EP 0610049 A **[0008]**
- EP 0780471 A **[0009]**
- US 08692002 B **[0107]**
- US 67711296 A **[0126] [0131]**
- WO 9116427 A **[0128]**
- US 65156896 A **[0131]**
- US 4946778 A **[0146]**
- US 60254414 B **[0161]**
- US 60173609 B **[0161]**

Non-patent literature cited in the description

- New family of carbon-nitrogen hydrolases. *Protein Science,* 1994, vol. 3, 1334-1348 **[0010]**
- *Gene,* 1995, vol. 161, 15-20 **[0011]**
- **T.E. CREIGHTON.** Proteins - Structure and Molecular Properties. W.H. Freeman and Company, 1993 **[0037]**
- Posttranslational Covalent Modification of Proteins. Academic Press, 1983, 1-12 **[0037]**
- **MERRIFIELD, R. B.** *J. Am. Chem. Soc.,* 1963, vol. 85, 2149-2154 **[0040]**
- **STEWART, J. M. ; YOUNG, J. D.** Solid Phase Peptide Synthesis. Pierce Chemical Co, 11-12 **[0040]**
- **H. M. GEYSEN et al.** *Proc. Natl. Acad. Sci., USA,* 1984, vol. 81, 3998 **[0040]**
- **B. LEWIN.** Genes VI. Oxford University Press, 1997, 214 **[0061]**
- **AUSUBEL et al.** Current Protocols in Molecular Biology. John Wiley 503 Sons, Inc, 1997 **[0068] [0110]**
- **SAMBROOK et al.** Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, 1989 **[0068] [0110]**
- **BARANY, F.** The Ligase Chain Reaction in a PCR World. *PCR Methods and Applications,* 1991, vol. 1, 5-16 **[0069]**
- **E. FAHY et al.** Self-sustained Sequence Replication (3SR): An Isothermal Transcription-based Amplification System Alternative to PCR. *PCR Methods and Applications,* 1991, vol. 1, 25-33 **[0069]**
- **WALKER G.T. et al.** Strand Displacement Amplification-an Isothermal in vitro DNA Amplification Technique. *Nucleic Acid Research,* 1992, vol. 20, 1691-1696 **[0069]**
- **SAMBROOK et al.** Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, 1989 **[0111]**
- **DAVIS, L. ; DIBNER, M. ; BATTEY, I.** *Basic Methods in Molecular Biology,* 1986 **[0114]**
- **GLUZMAN.** *Cell,* 1981, vol. 23, 175 **[0117]**
- **LEUNG, D.W. et al.** *Technique,* 1989, vol. 1, 11-15 **[0124]**
- **CALDWELL, R. C. ; JOYCE G.F.** *PCR Methods Applic.,* 1992, vol. 2, 28-33 **[0124]**
- **REIDHAAR-OLSON, J.F. ; SAUER, R.T. et al.** *Science,* 1988, vol. 241, 53-57 **[0125]**
- **STEMMER, W.P.** *PNAS, USA,* 1994, vol. 91, 10747-10751 **[0127]**
- **ARKIN, A.P. ; YOUVAN, D.C.** *PNAS, USA,* 1992, vol. 89, 7811-7815 **[0130]**
- **DELEGRAVE, S. ; YOUVAN, D.C.** *Biotechnology Research,* 1993, vol. 11, 1548-1552 **[0131]**
- **ARNOLD, F.H.** *Current Opinion in Biotechnology,* 1993, vol. 4, 450-455 **[0131]**
- **KOHLER ; MILSTEIN.** *Nature,* 1975, vol. 256, 495-497 **[0145]**
- **KOZBOR et al.** *Immunology Today,* 1983, vol. 4, 72 **[0145]**
- **COLE et al.** Monoclonal Antibodies and Cancer Therapy. Alan R. Liss, Inc, 1985, 77-96 **[0145]**
- Methods for Measuring Cellulase Activities. *Methods in Enzymology,* vol. 160, 87-116 **[0147]**
- **STRYER, LUBERT.** Biochemistry. W. H Freeman & Co, **[0148]**
- **STARRIER, LUBERT.** Biochemistry. W. H Freeman & Co, **[0149]**
- **PEARSON ; LIPMAN.** *Proc. Natl. Acad. Sci. USA,* 1988, vol. 85 (8), 2444-2448 **[0152]**
- **ALTSCHUL et al.** *J. Mol. Biol.,* 1990, vol. 215 (3), 403-410 **[0152]**
- **THOMPSON et al.** *Nucleic Acids Res.,* 1994, vol. 22 (2), 4673-4680 **[0152]**
- **HIGGINS et al.** *Methods Enzymol.,* 1996, vol. 266, 383-402 **[0152]**
- **ALTSCHUL et al.** *Nature Genetics,* 1993, vol. 3, 266-272 **[0152]**
- **SMITH ; WATERMAN.** *Adv. Appl. Math.,* 1981, vol. 2, 482 **[0152]**
- **NEEDLEMAN ; WUNSCH.** *J. Mol. Biol,* 1970, vol. 48, 443 **[0152]**

- **PERSON ; LIPMAN.** *Proc. Nat'l. Acad. Sci. USA,* 1988, vol. 85, 2444 **[0152]**
- **ALTSCHUL et al.** *Nuc. Acids Res.,* 1977, vol. 25, 3389-3402 **[0152]**
- **ALTSCHUL et al.** *J. Mol. Biol.,* 1990, vol. 215, 403-410 **[0152]**
- **HENIKOFF ; HENIKOFF.** *Proc. Natl. Acad. Sci. USA,* 1989, vol. 89, 10915 **[0152]**
- **KARLIN ; ALTSCHUL.** *Proc. Natl. Acad. Sci. USA,* 1993, vol. 90, 5873 **[0152]**
- **GONNET et al.** *Science,* 1992, vol. 256, 1443-1445 **[0152]**
- **HENIKOFF ; HENIKOFF.** *Proteins,* 1993, vol. 17, 49-61 **[0152]**
- Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, 1978 **[0152]**
- **ALTSCHUL et al.** *J. Mol. Biol.,* 1990, vol. 215, 403 **[0156]**
- **PEARSON ; LIPMAN.** *Proc. Natl. Acad. Sci. USA,* 1988, vol. 85, 2444 **[0156]**
- **BRUTLAG et al.** *Comp. App. Biosci.,* 1990, vol. 6, 237-245 **[0156]**